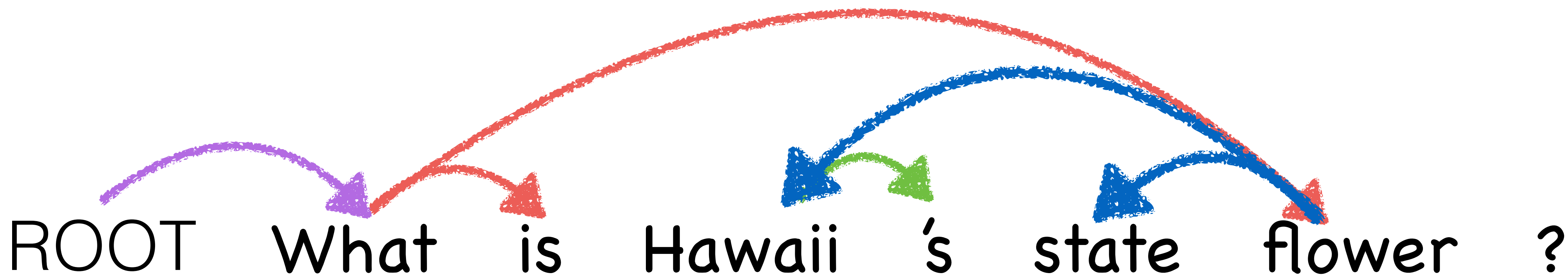
Dependency-based Convolutional Neural Networks for Sentence Embedding



Mingbo Ma Liang Huang Bing Xiang Bowen Zhou

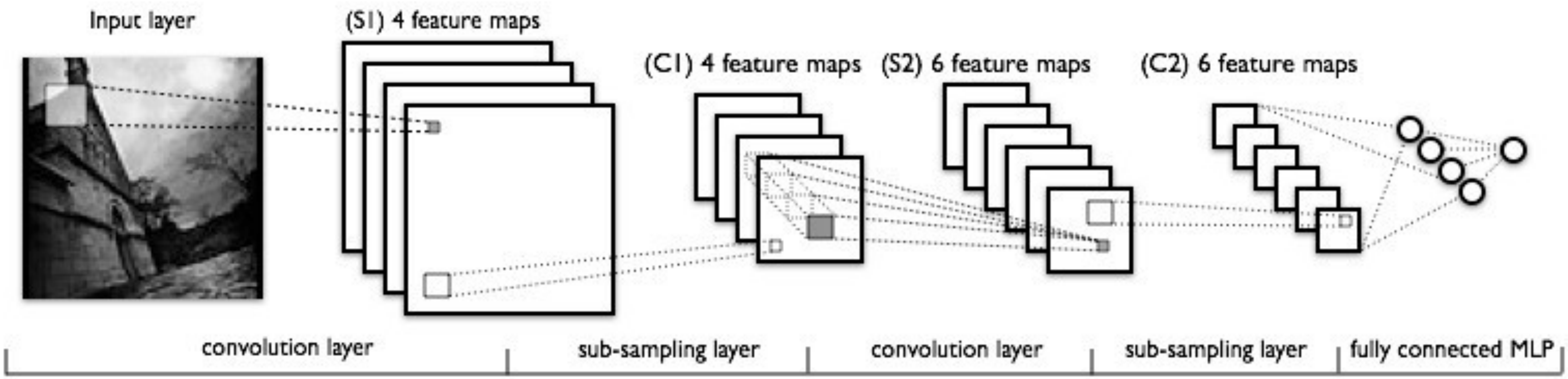
CUNY **IBM T. J. Watson**

The City University of New York

IBM

ACL 2015
Beijing

Convolutional Neural Network for NLP



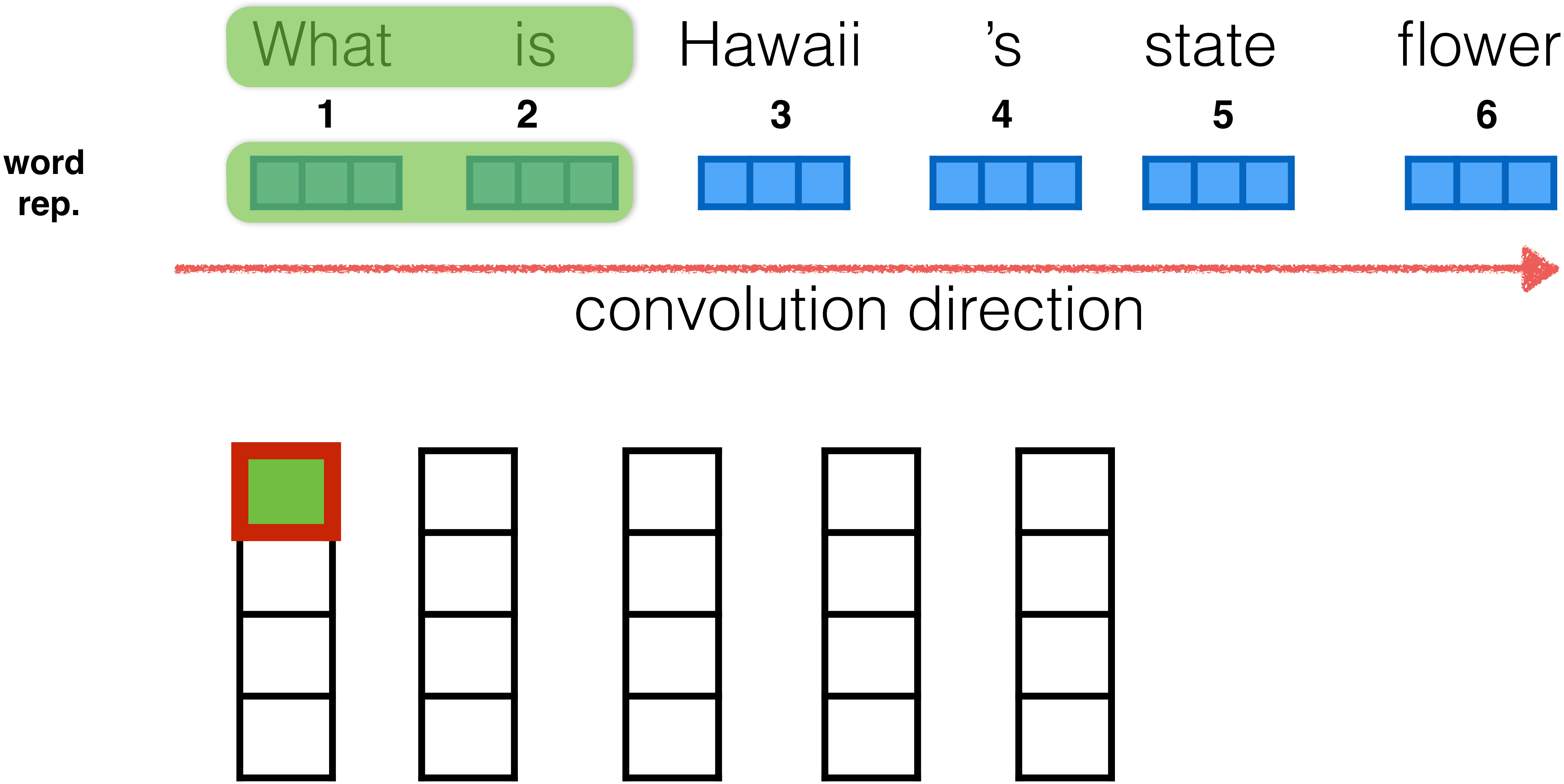
Kalchbrenner et al. (2014) and Kim (2014) apply CNNs to sentence modeling

- alleviates data sparsity by word embedding
- sequential order (sentence) instead of spatial order (image)

Should use more linguistic and structural information!

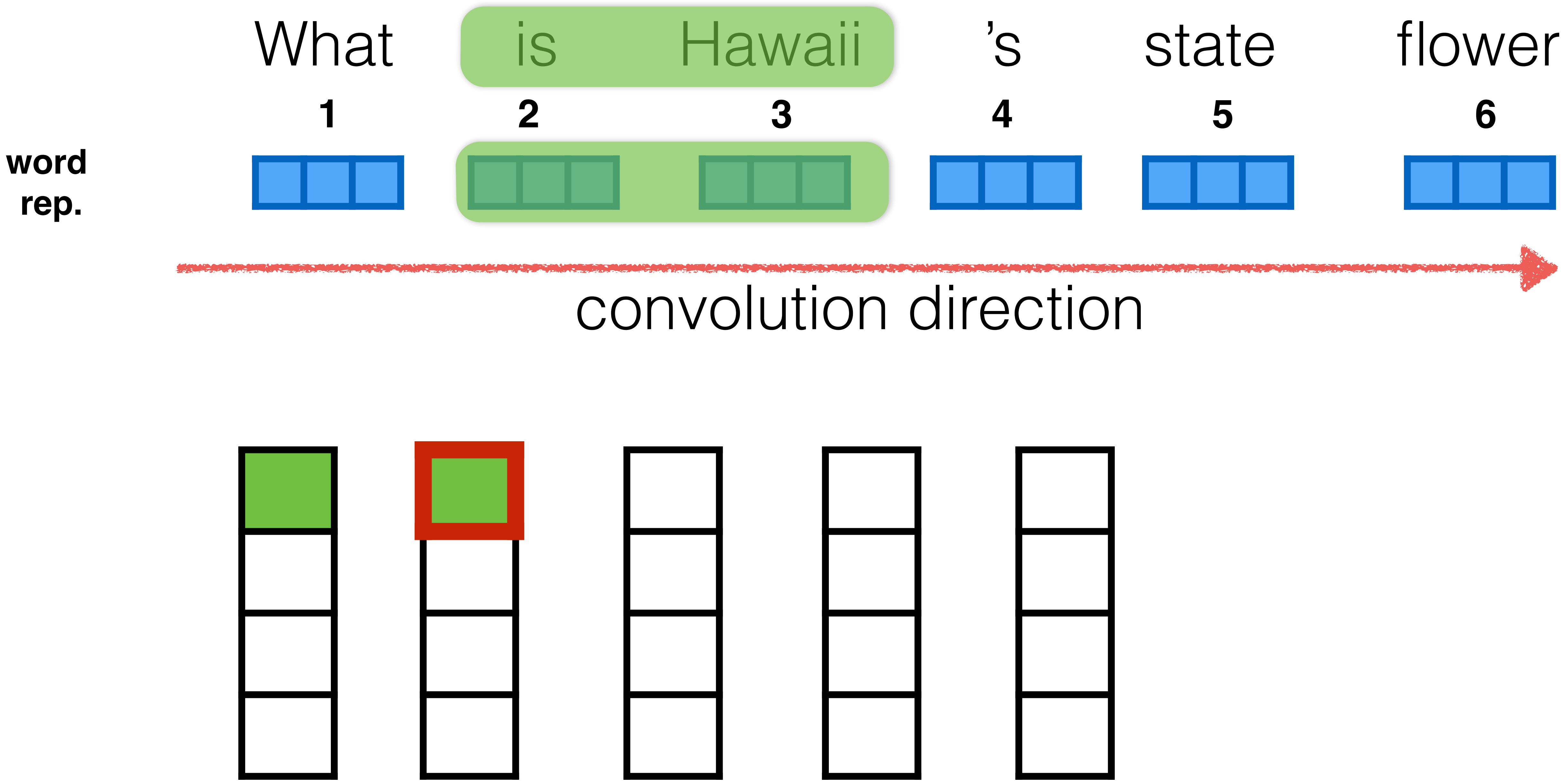
Sequential Convolution

Sequential convolution



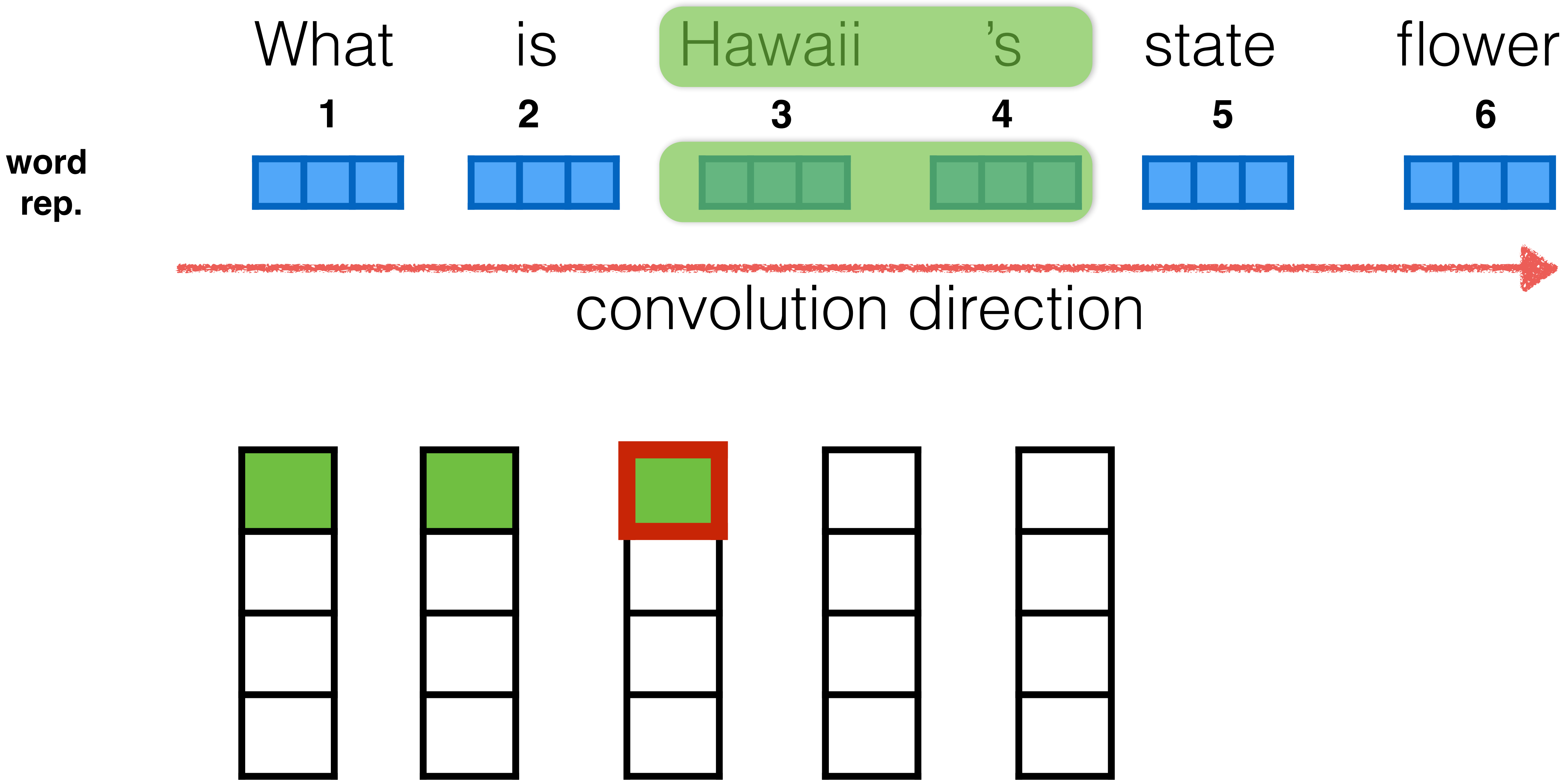
Sequential Convolution

Sequential convolution



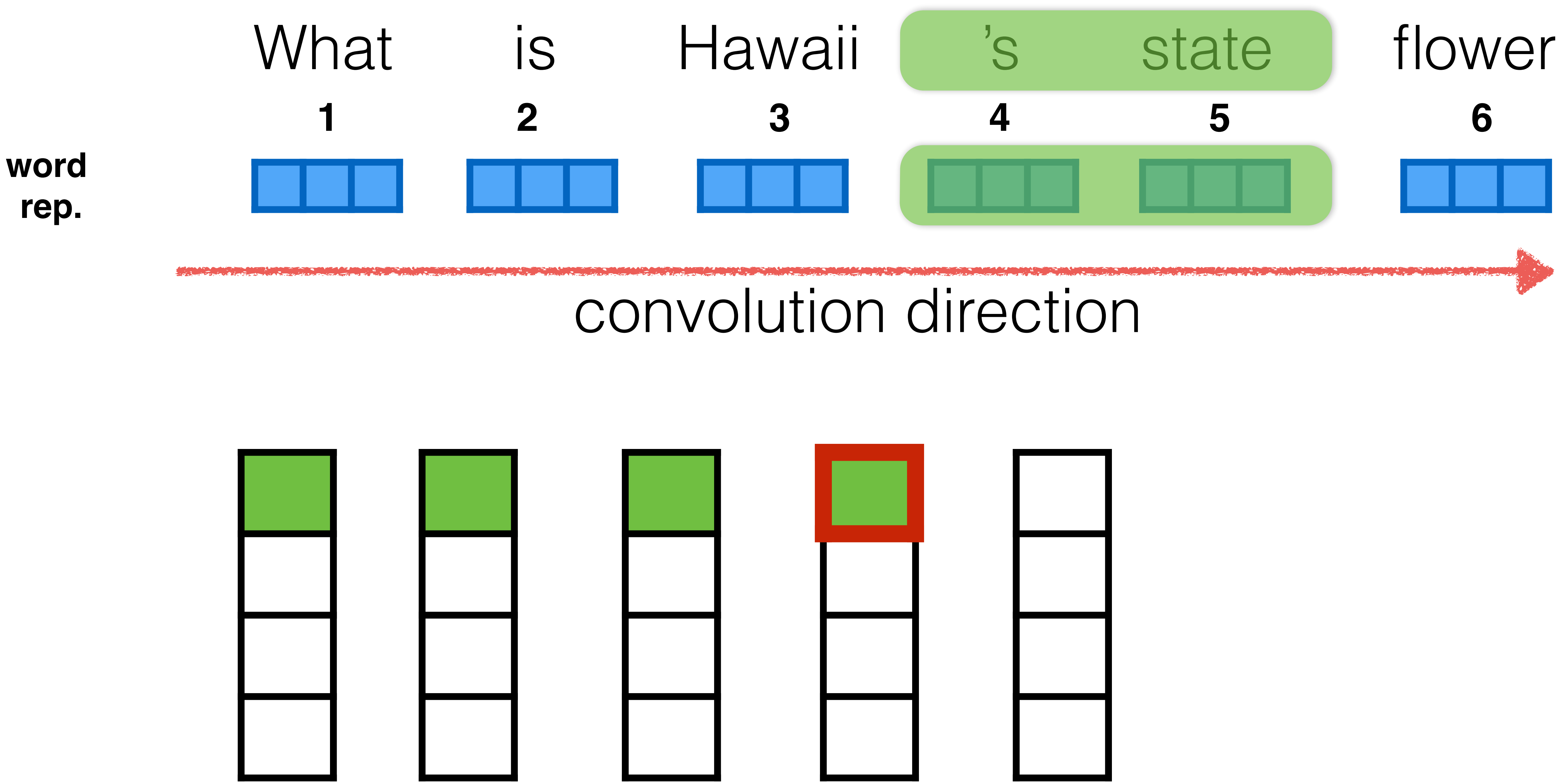
Sequential Convolution

Sequential convolution



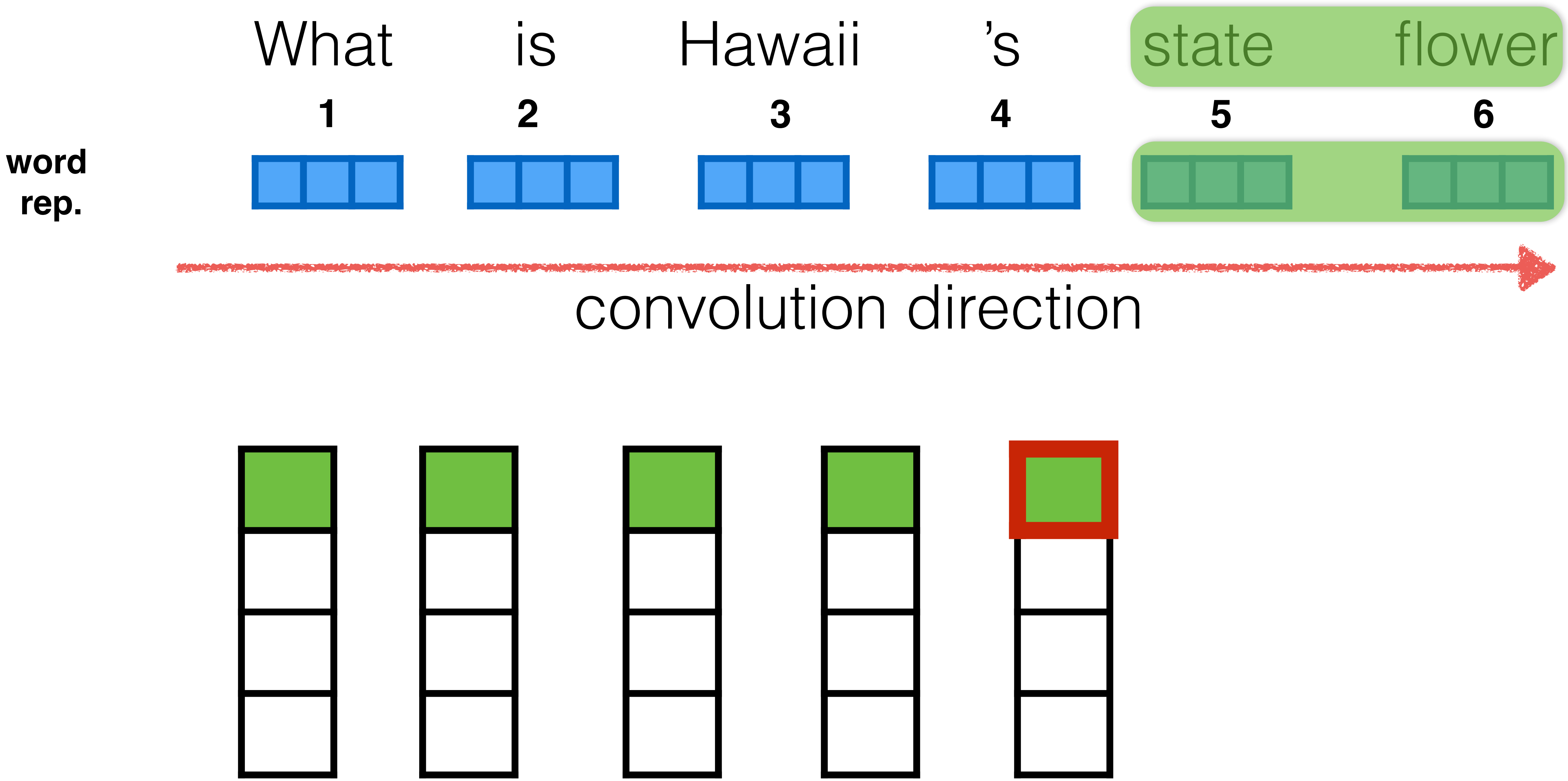
Sequential Convolution

Sequential convolution



Sequential Convolution

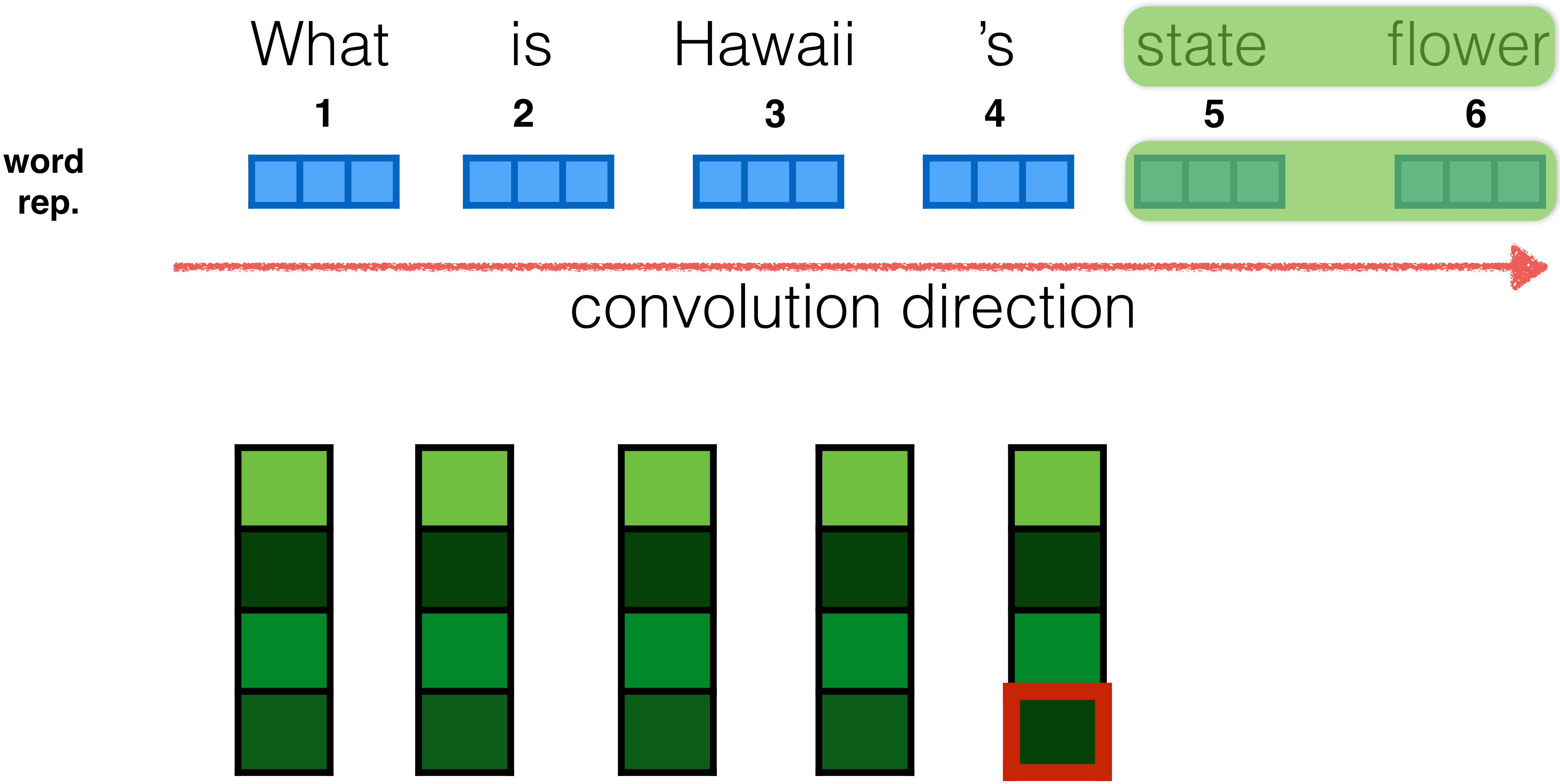
Sequential convolution



Try different convolution filters and repeat the same process

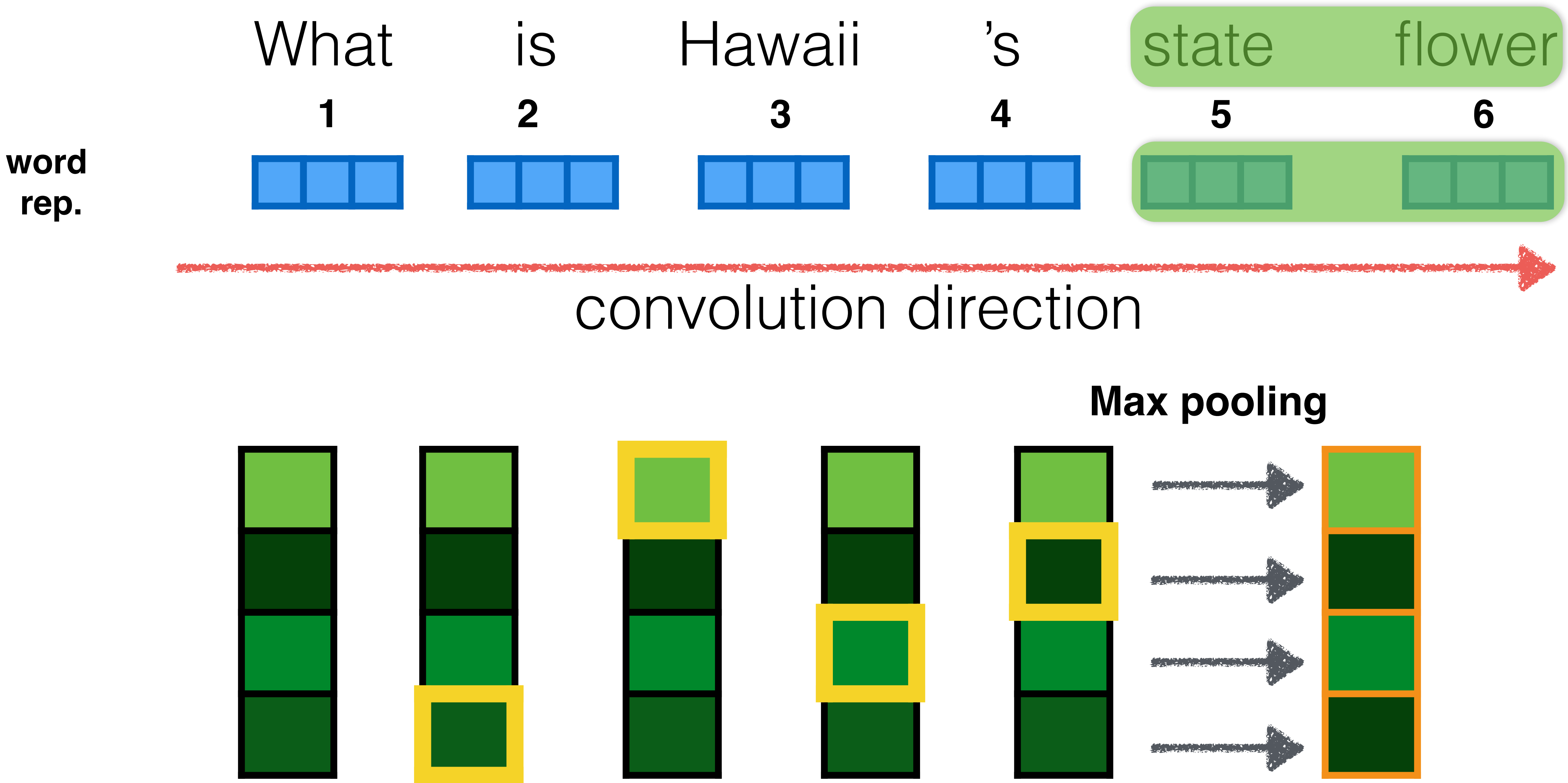
Sequential Convolution

Sequential convolution



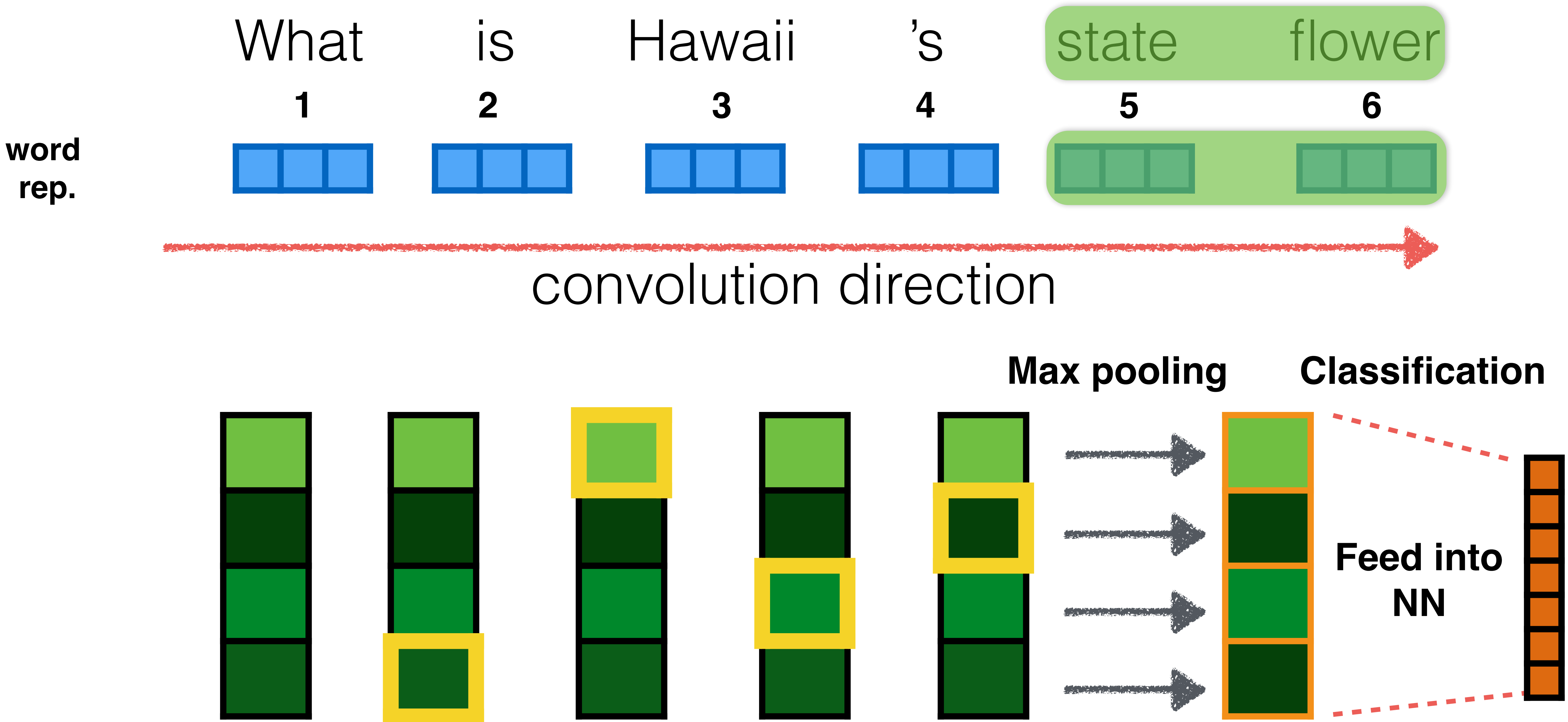
Sequential Convolution

Sequential convolution



Sequential Convolution

Sequential convolution



Example: Question Type Classification (TREC)

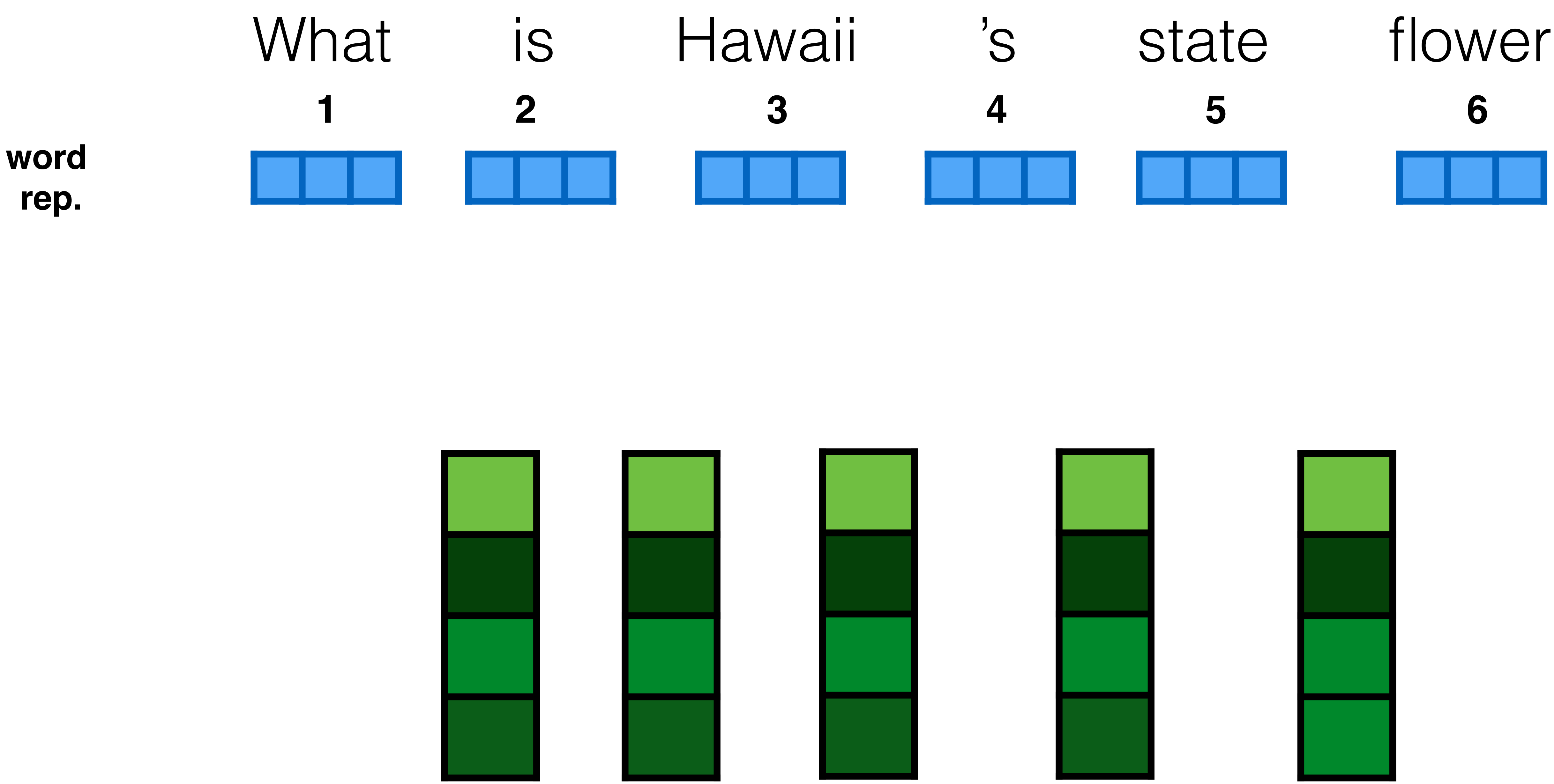
Sequential Convolution: Location

What is Hawaii 's state flower ?

Gold standard: Entity

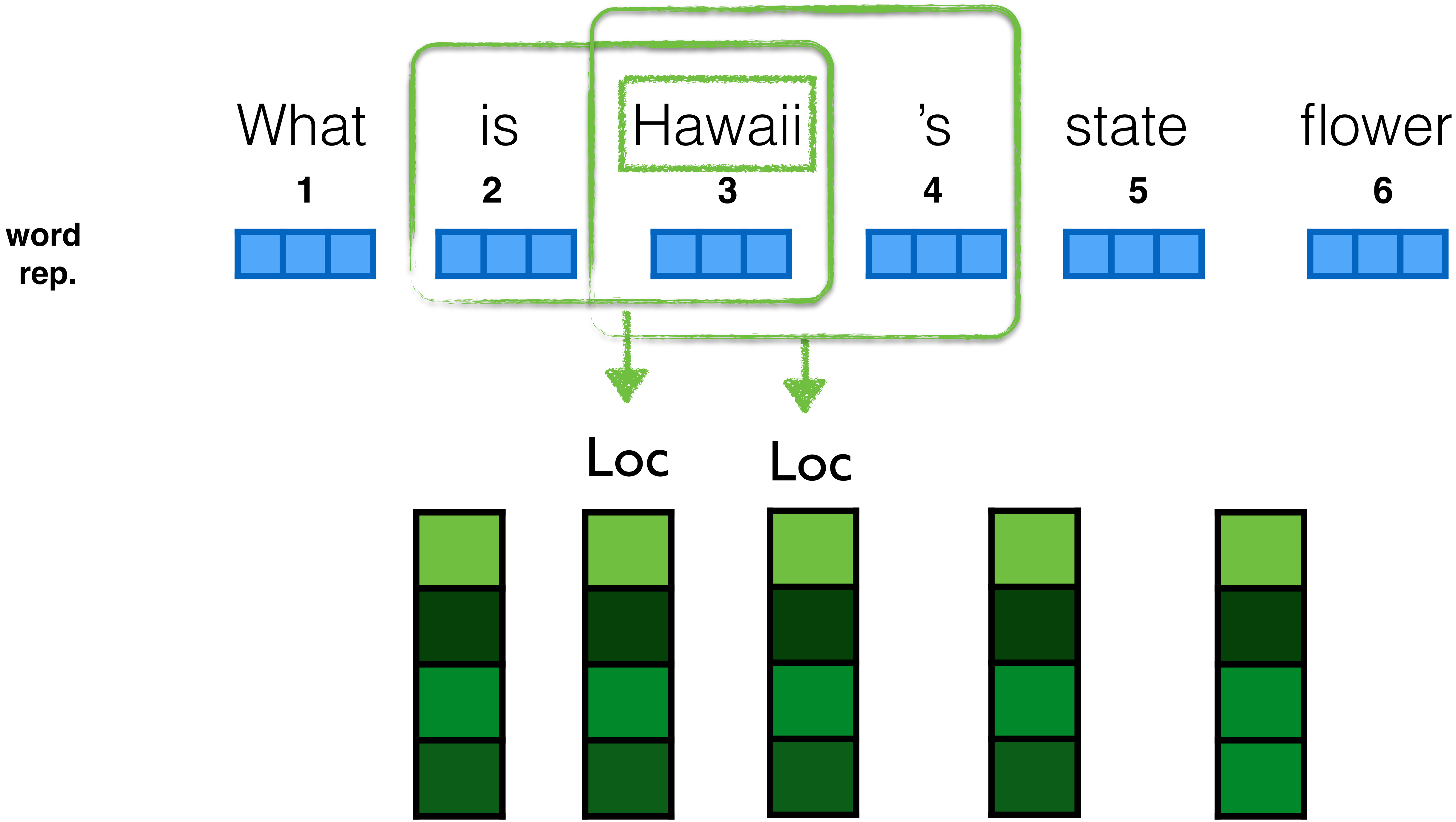
Sequential Convolution

Sequential convolution



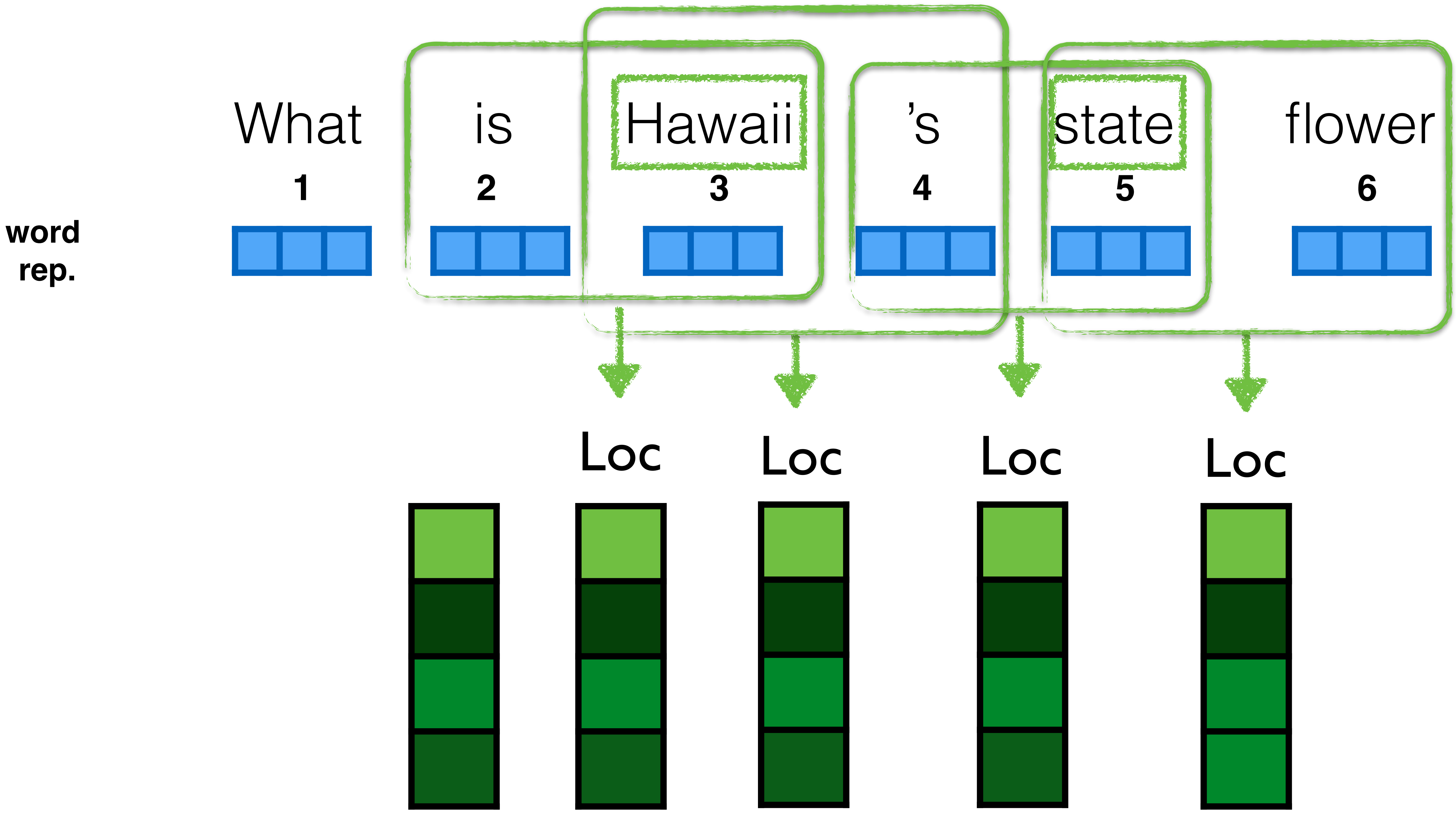
Sequential Convolution

Sequential convolution



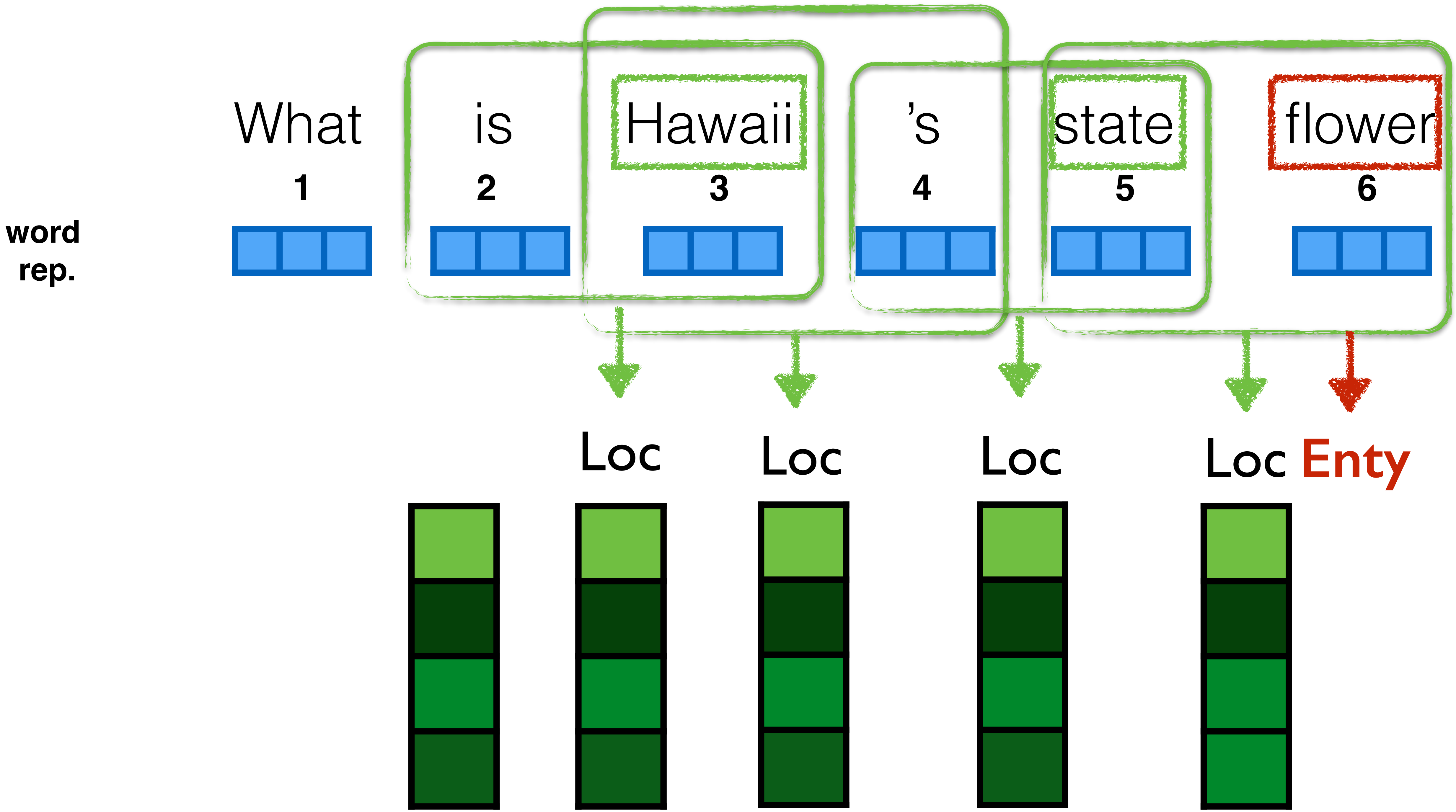
Sequential Convolution

Sequential convolution



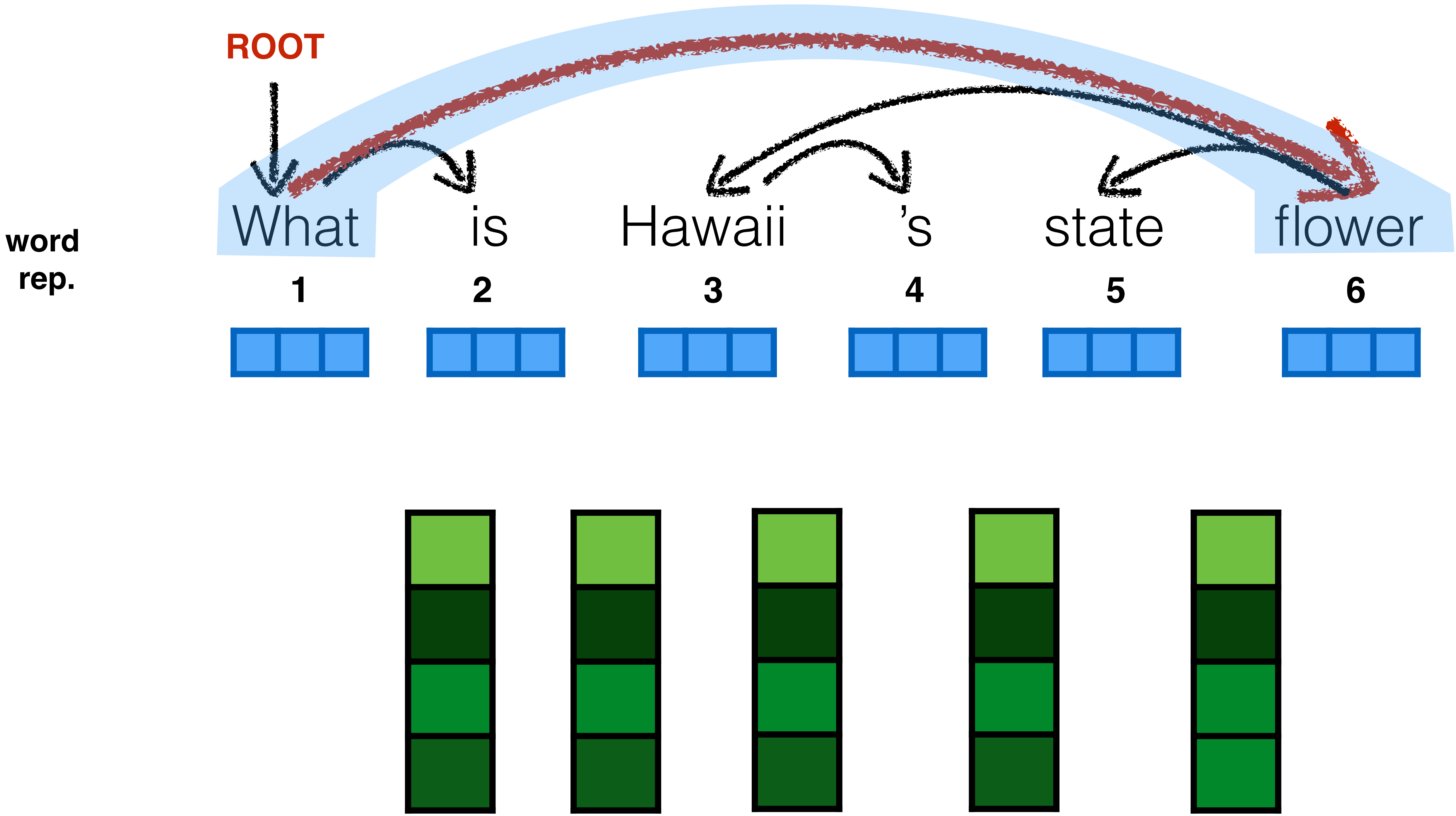
Sequential Convolution

Sequential convolution



Convolution on Tree

Sequential convolution



Sequential Convolution

Sequential convolution:

- Traditional convolution operates in surface order
- Cons: No structural information is captured

  No long distance relationships

Dependency-based Convolution

Sequential convolution:

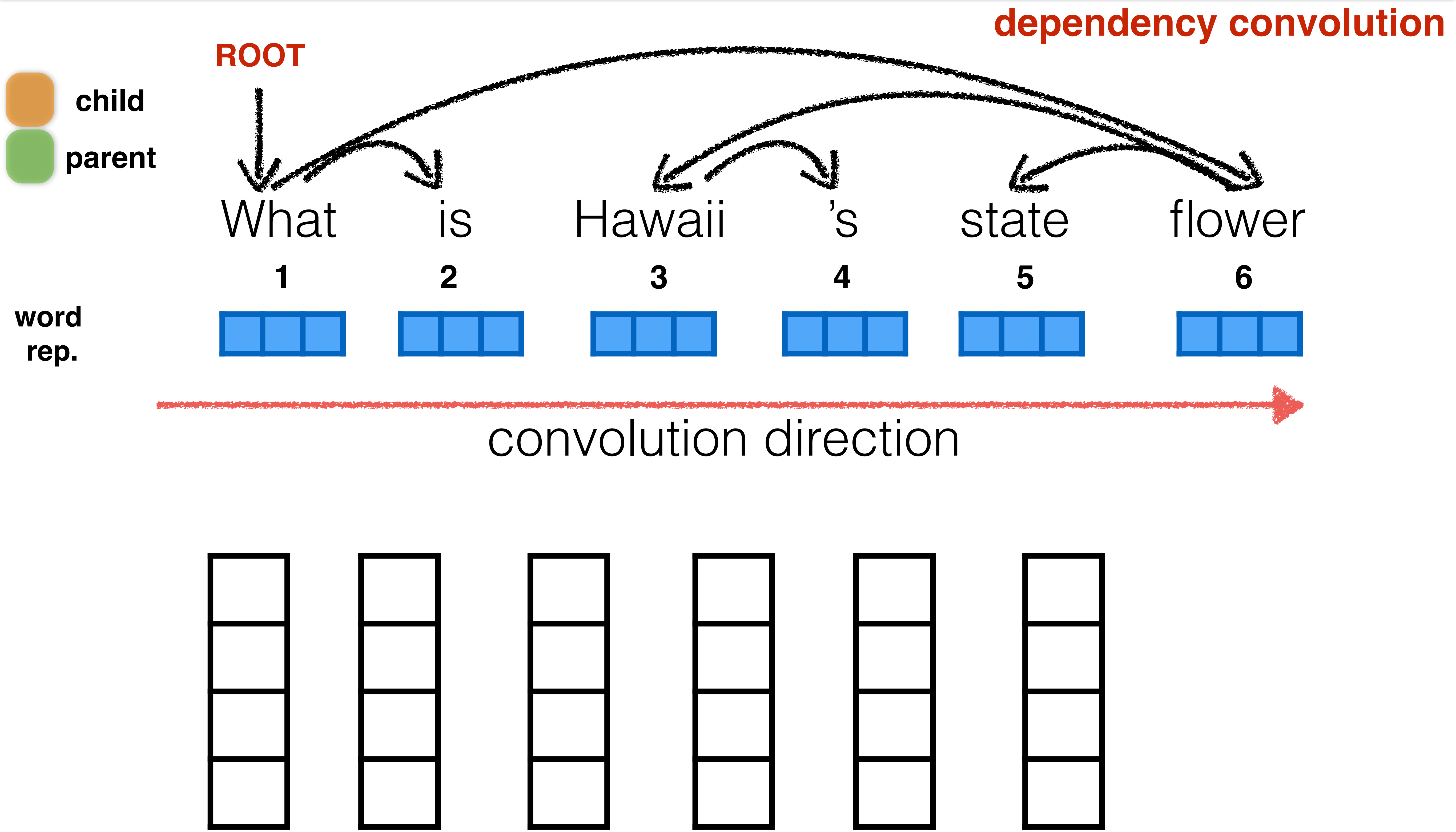
- Traditional convolution operates in surface order
- Cons: No structural information is captured

  No long distance relationships

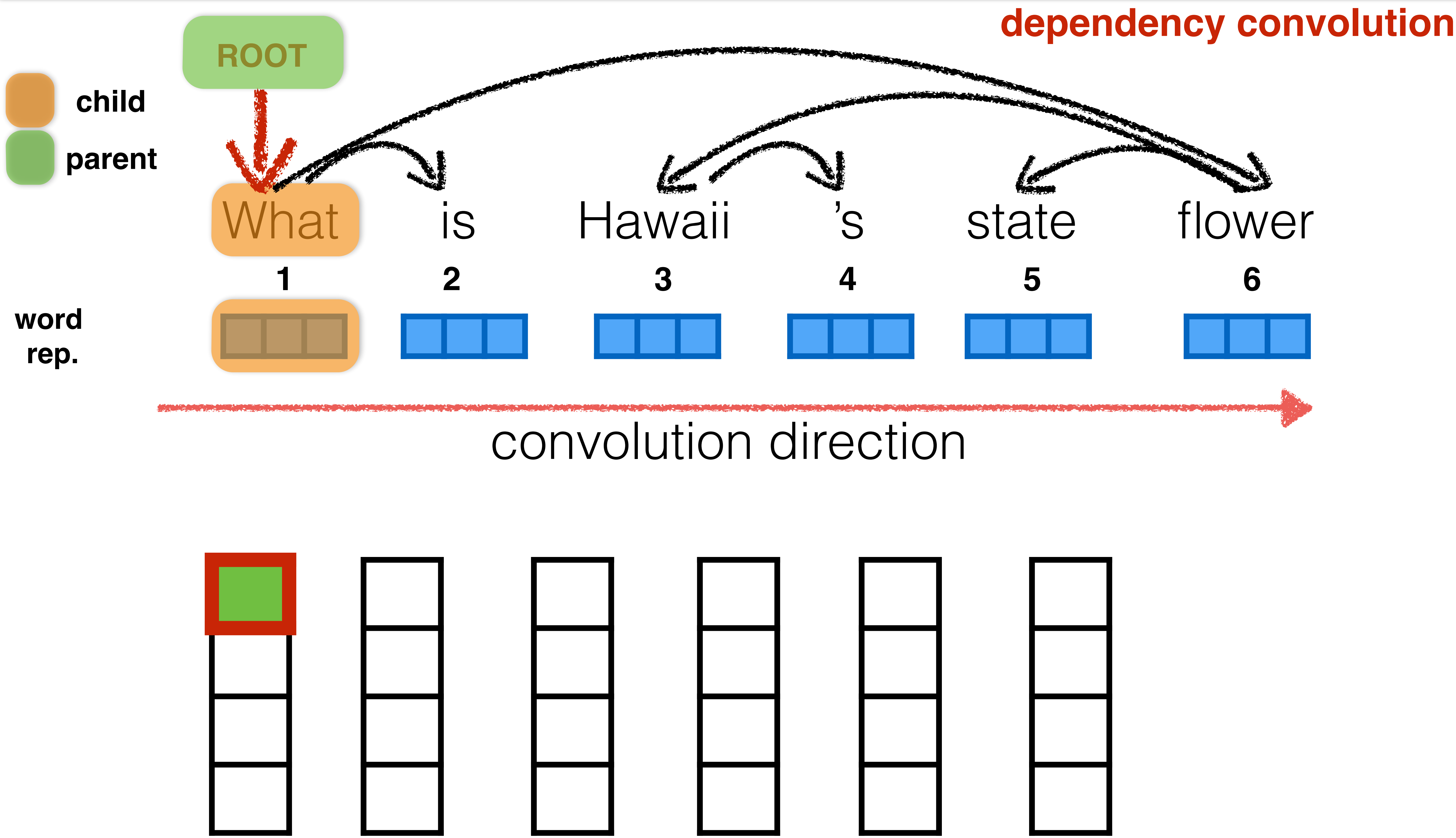
Structural Convolution:

- operates the convolution filters on dependency tree
- more "important" words are convolved more often
- long distance relationships is naturally obtained

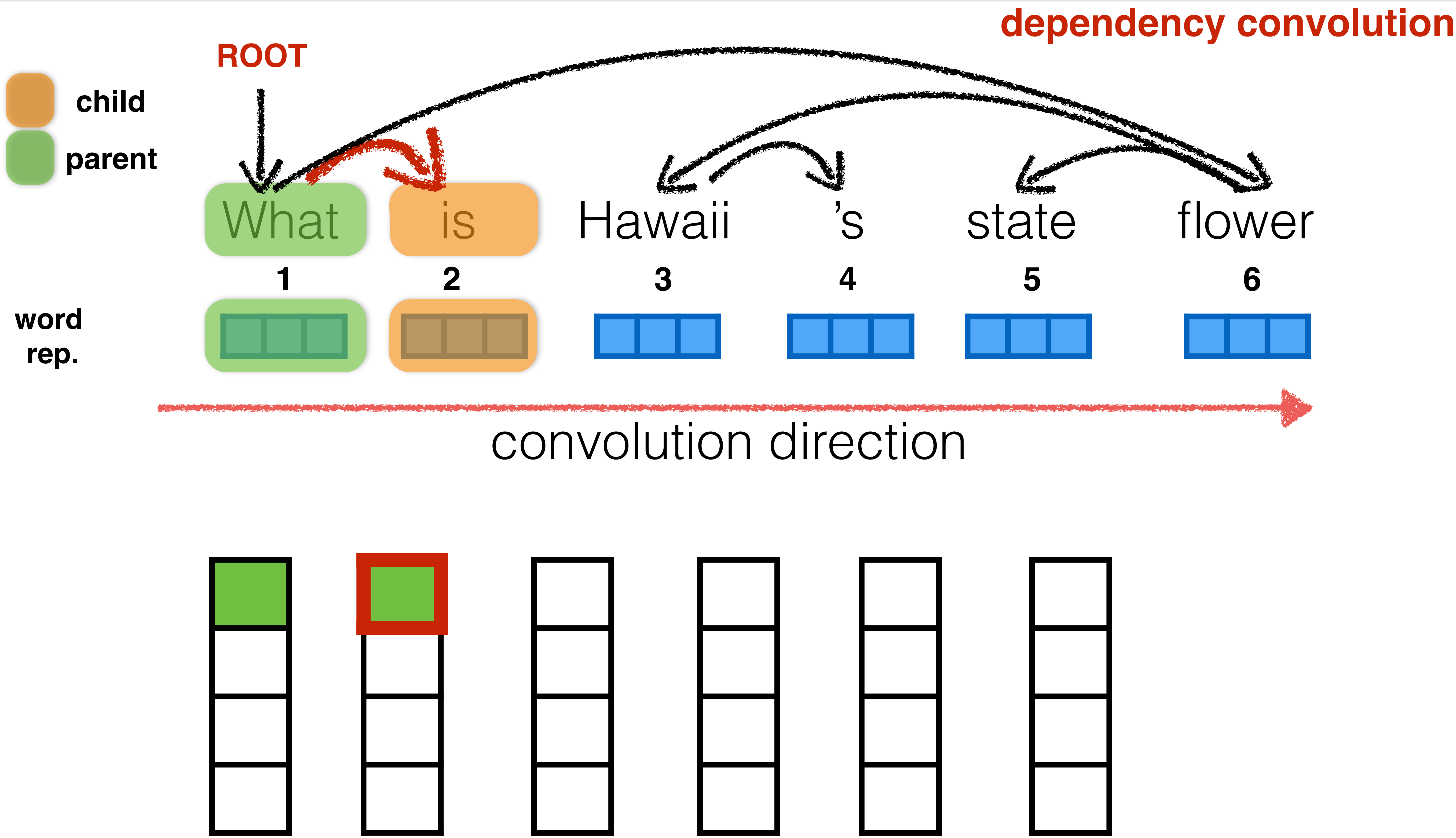
Convolution on Tree



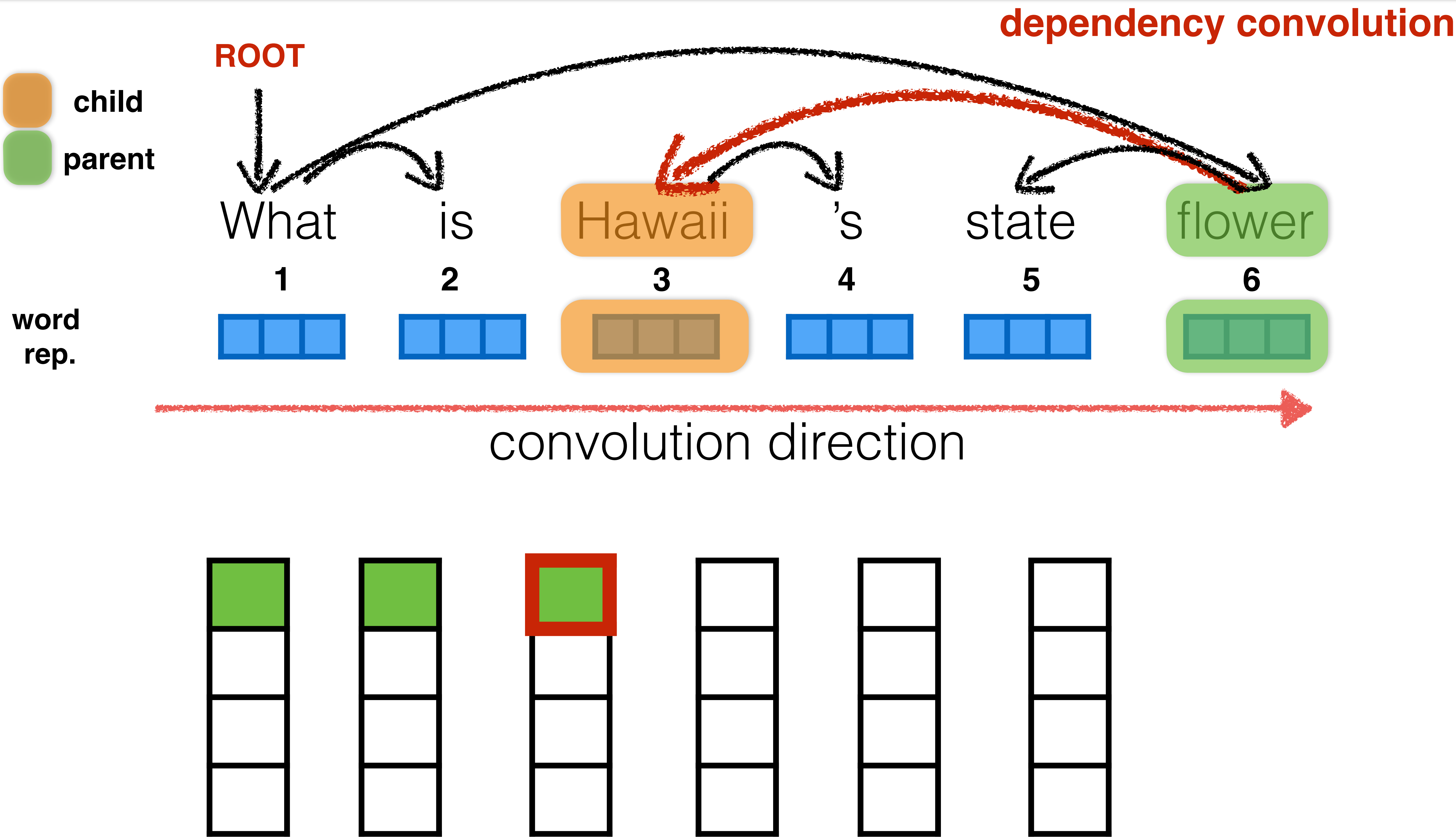
Convolution on Tree



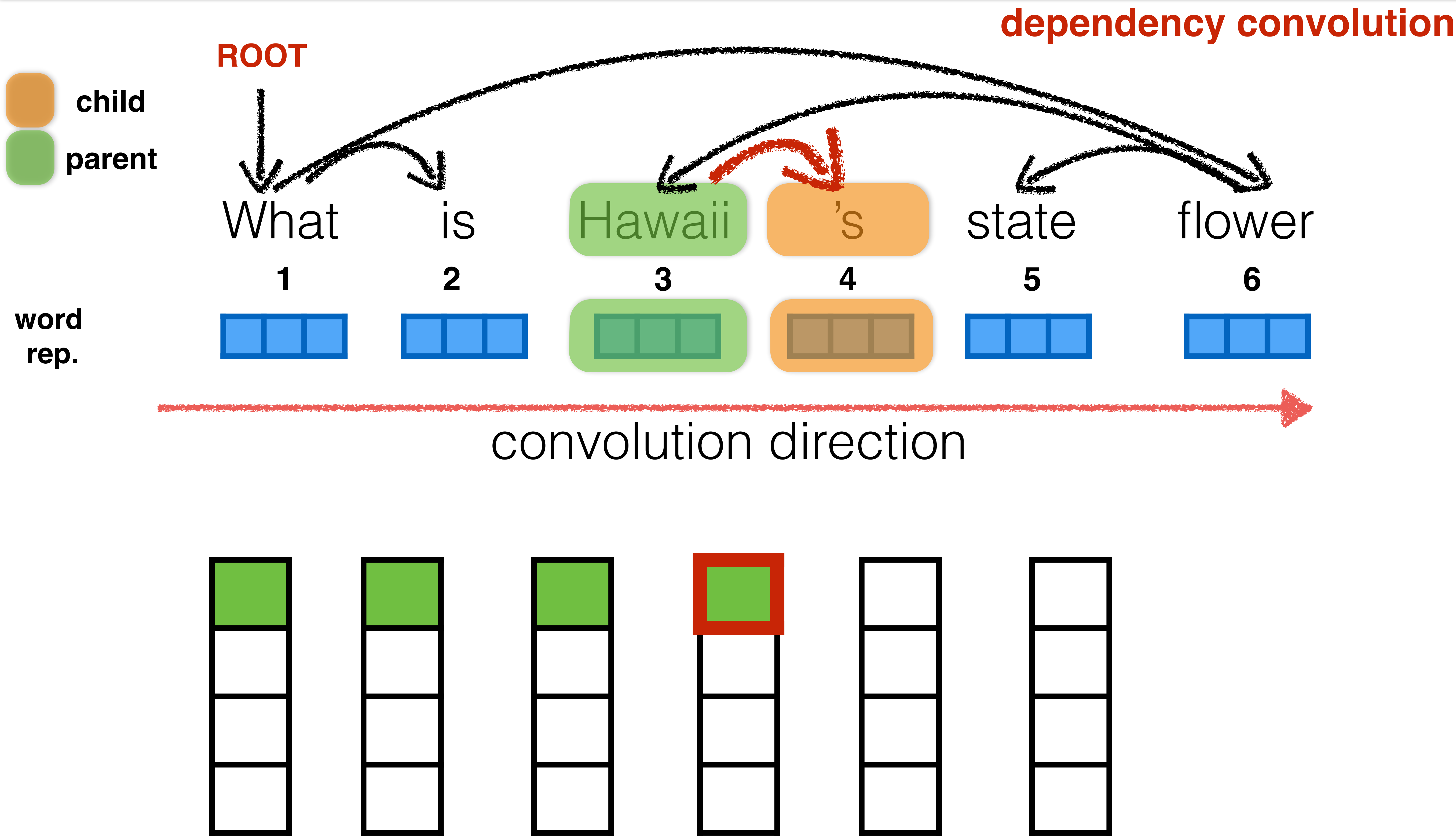
Convolution on Tree



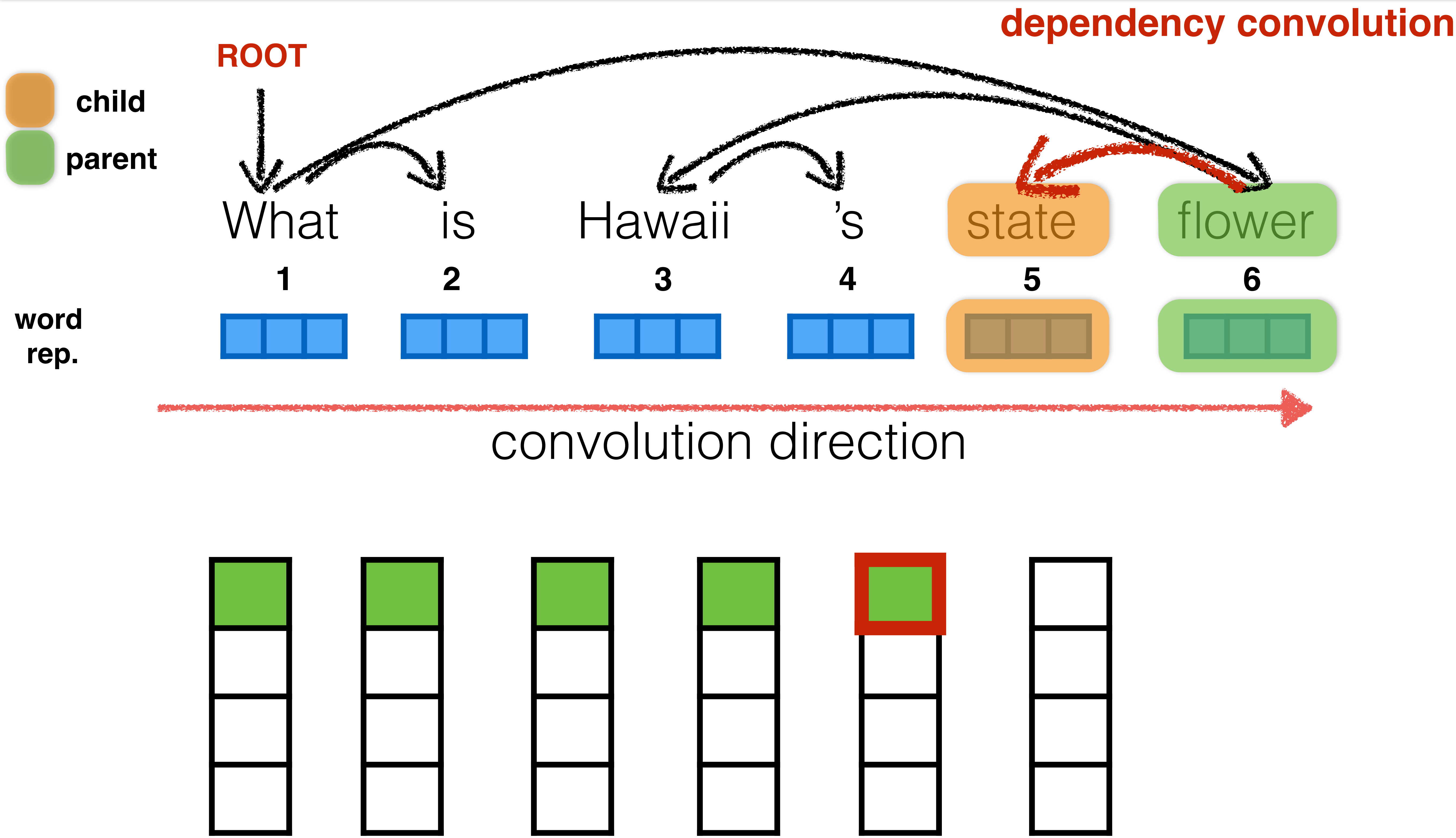
Convolution on Tree



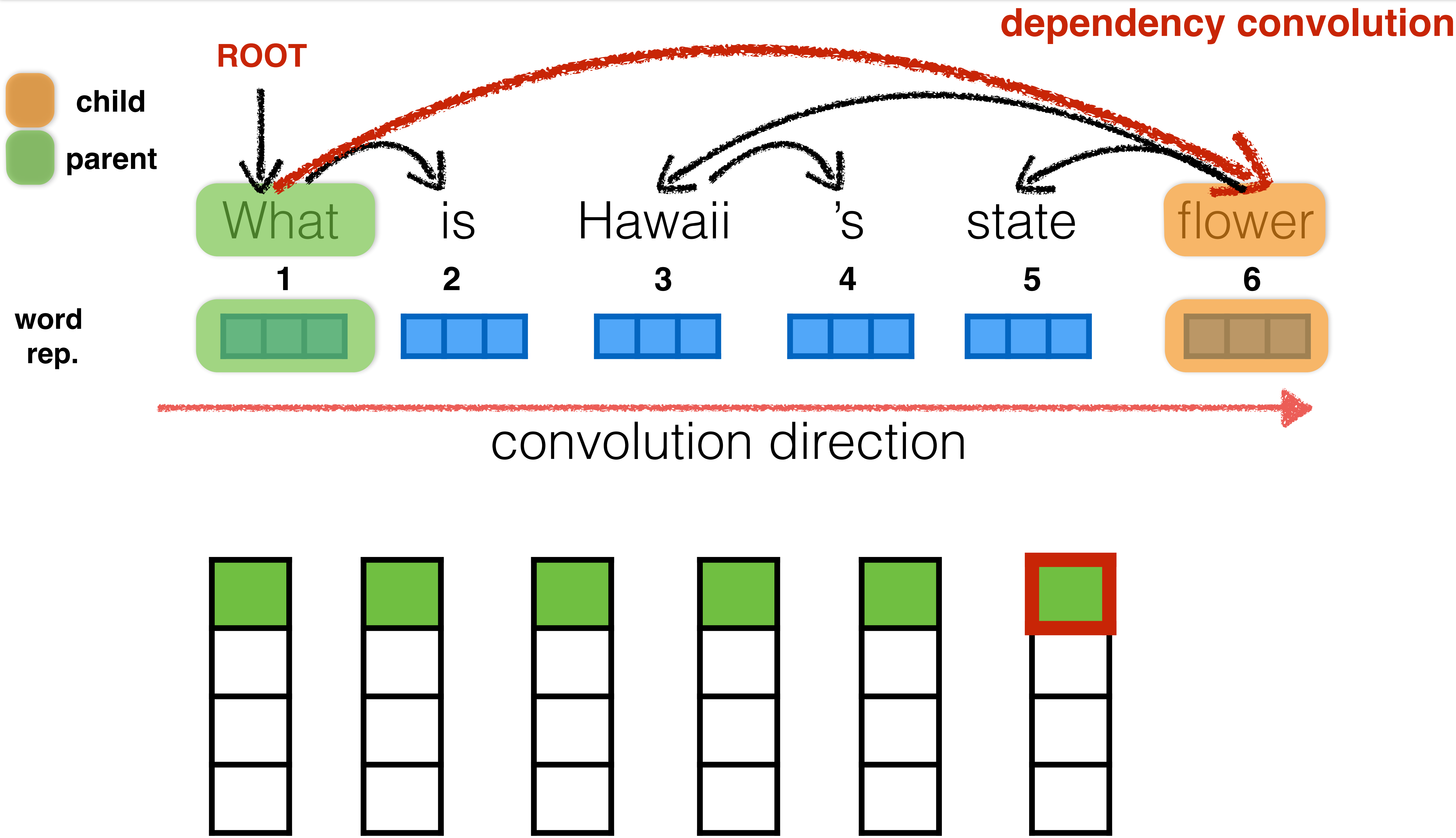
Convolution on Tree



Convolution on Tree

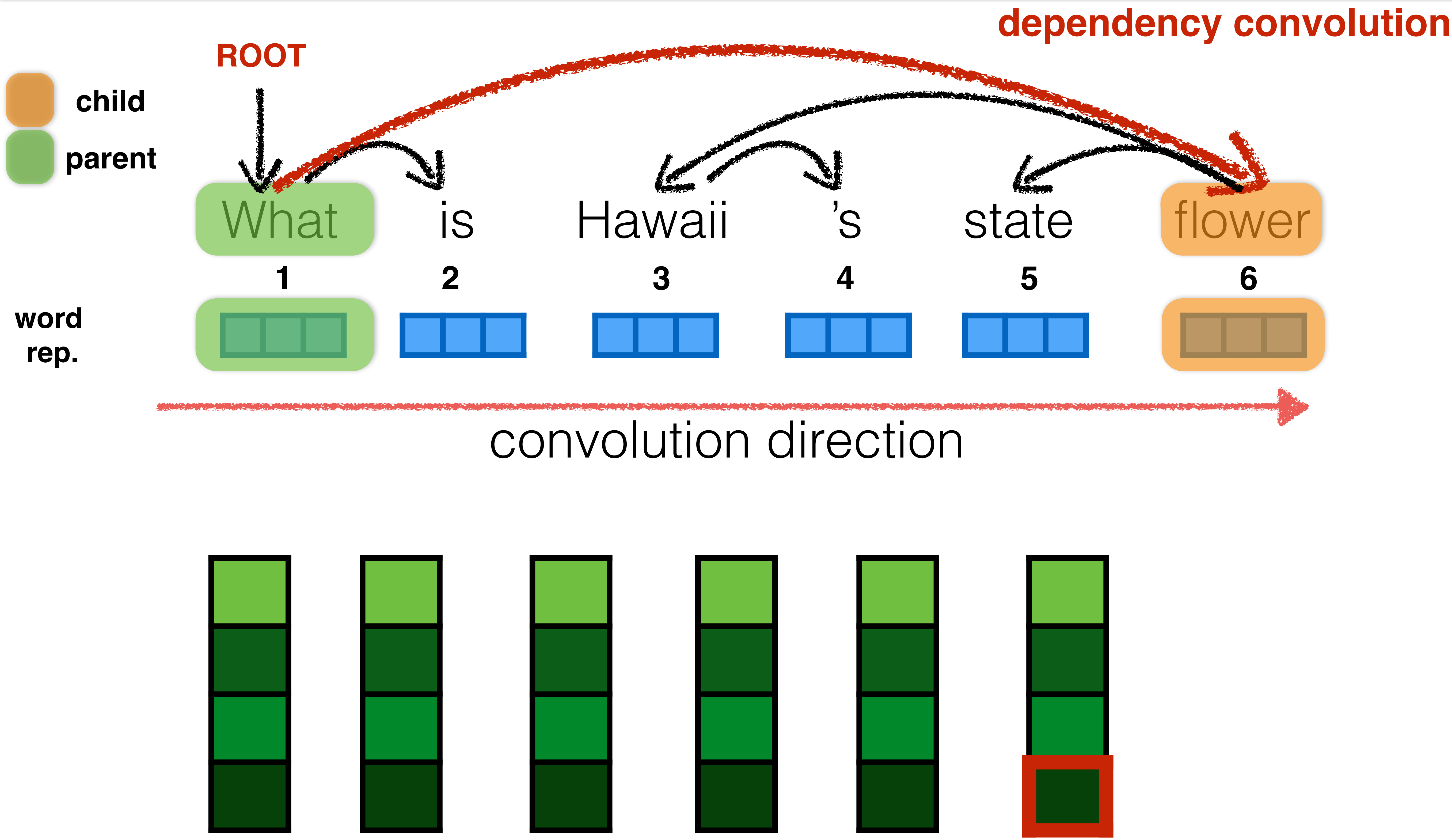


Convolution on Tree

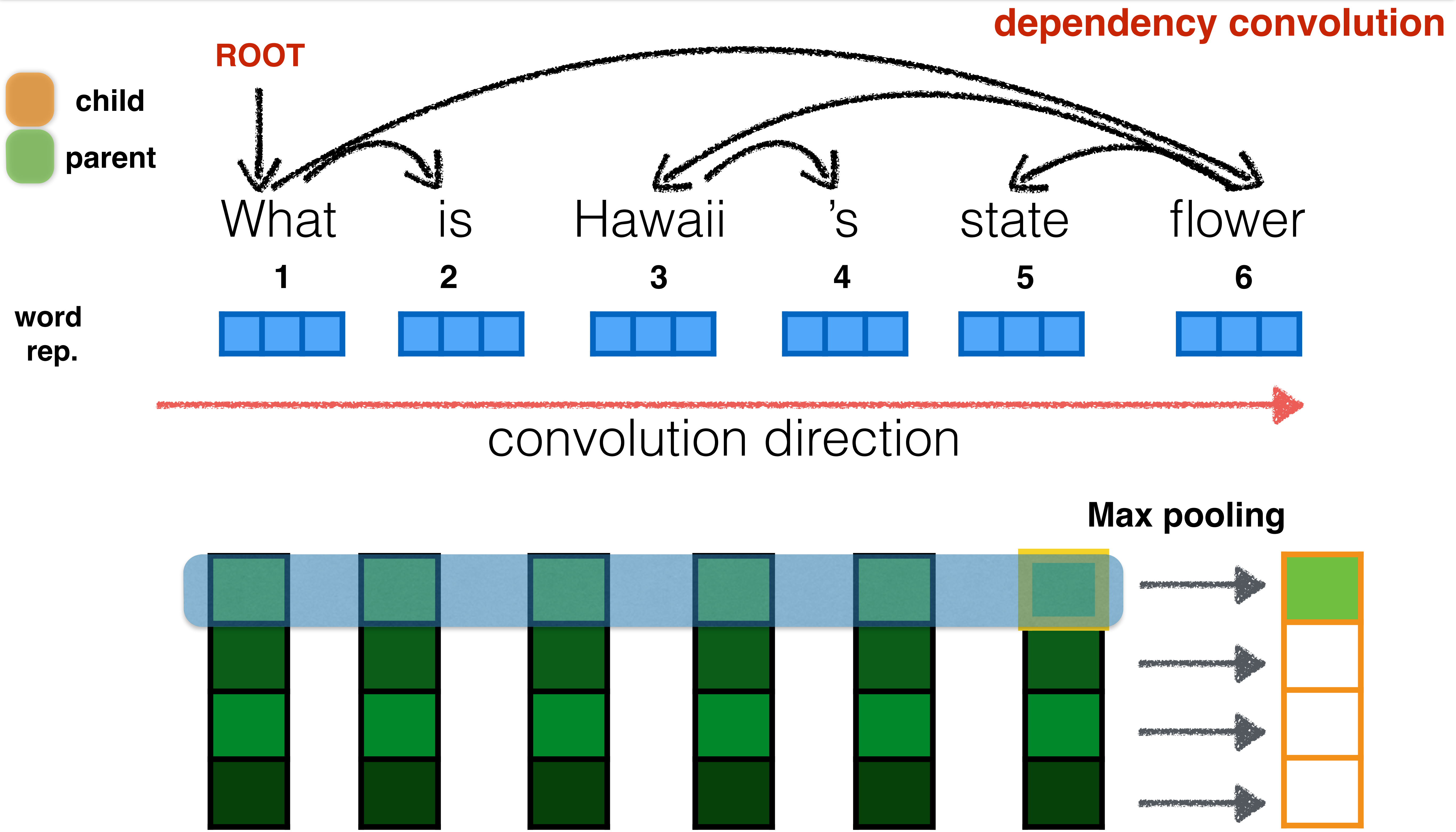


Try different **Bigram** convolution filters and repeat the same process

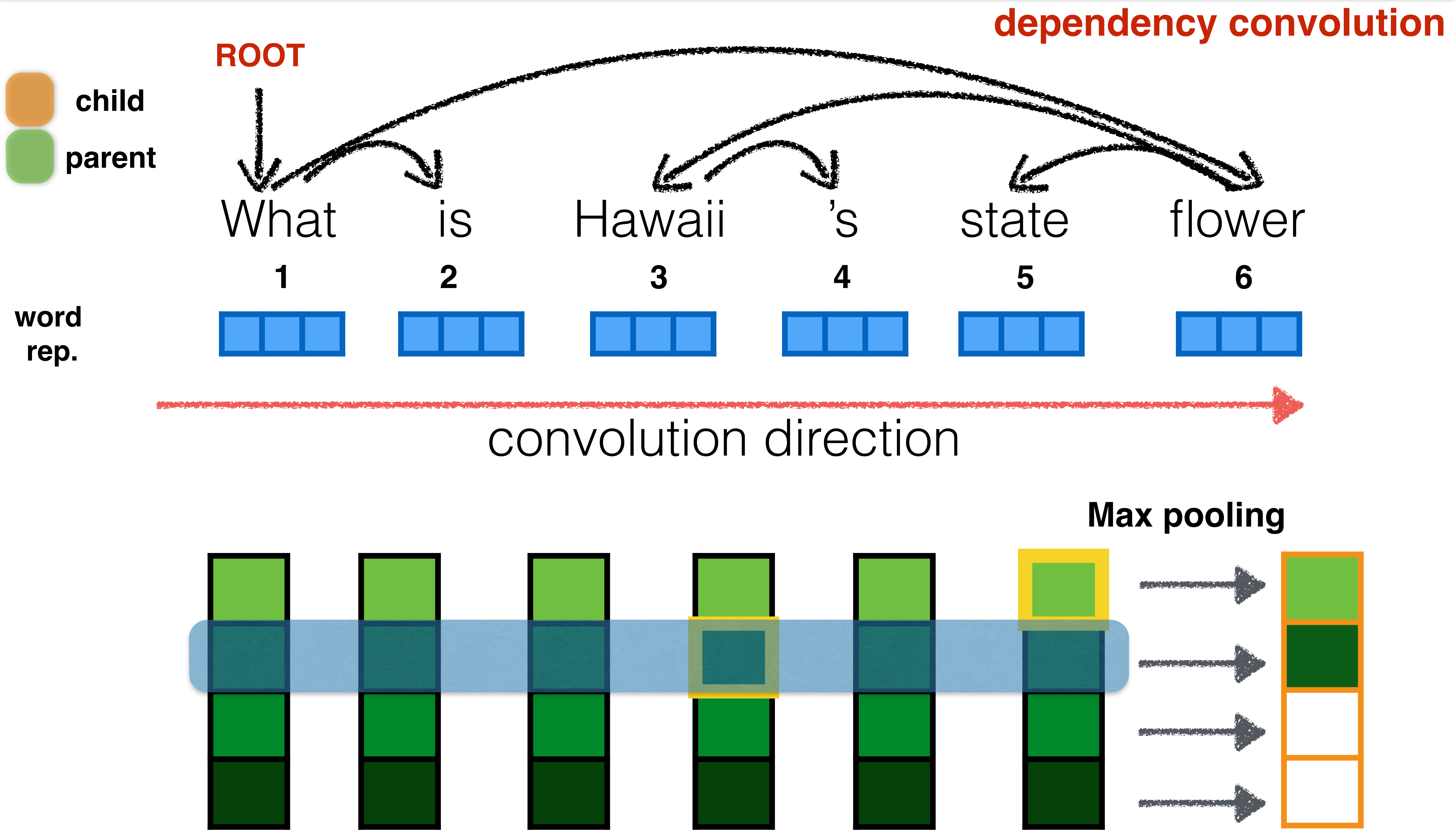
Convolution on Tree



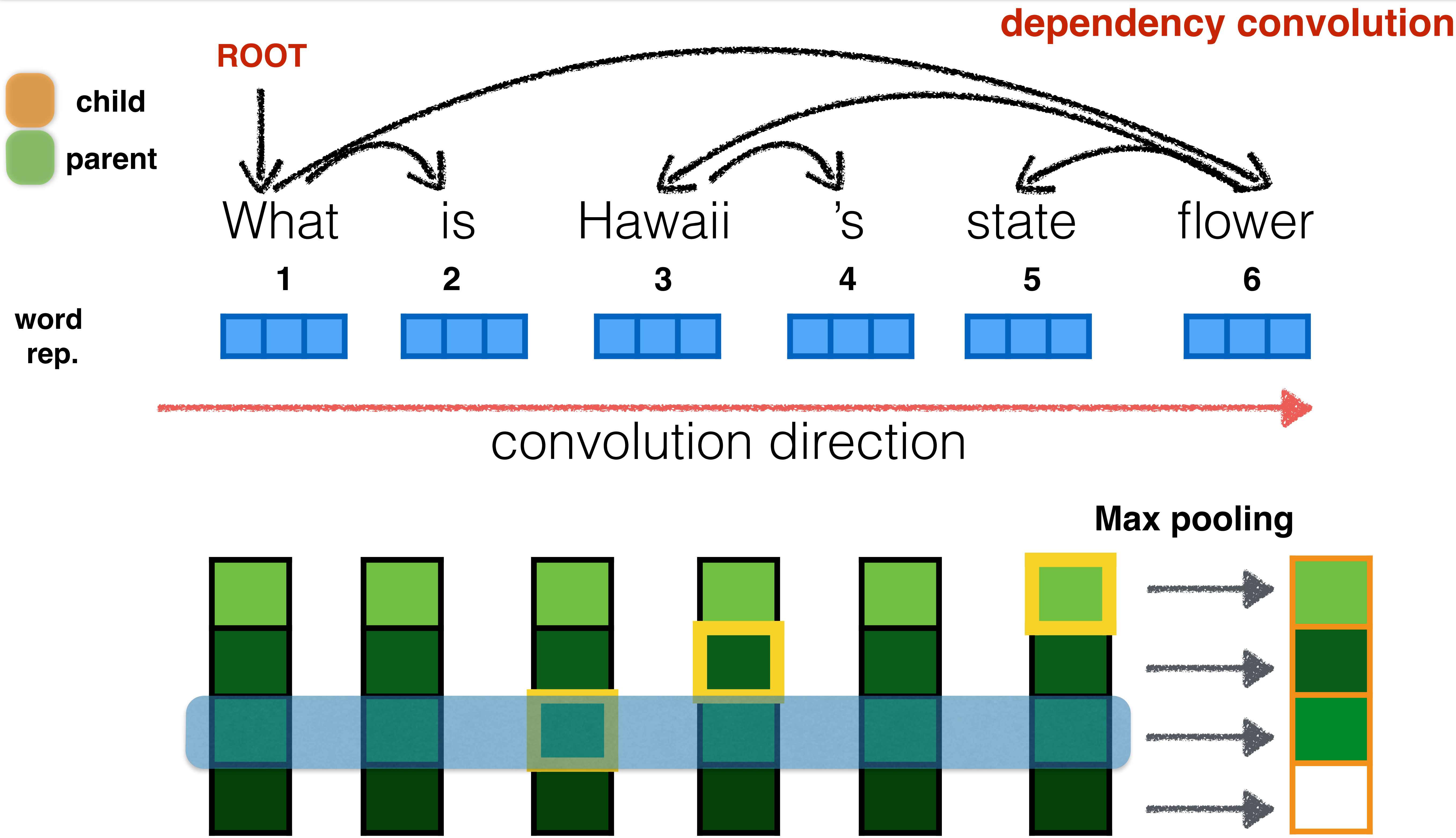
Convolution on Tree



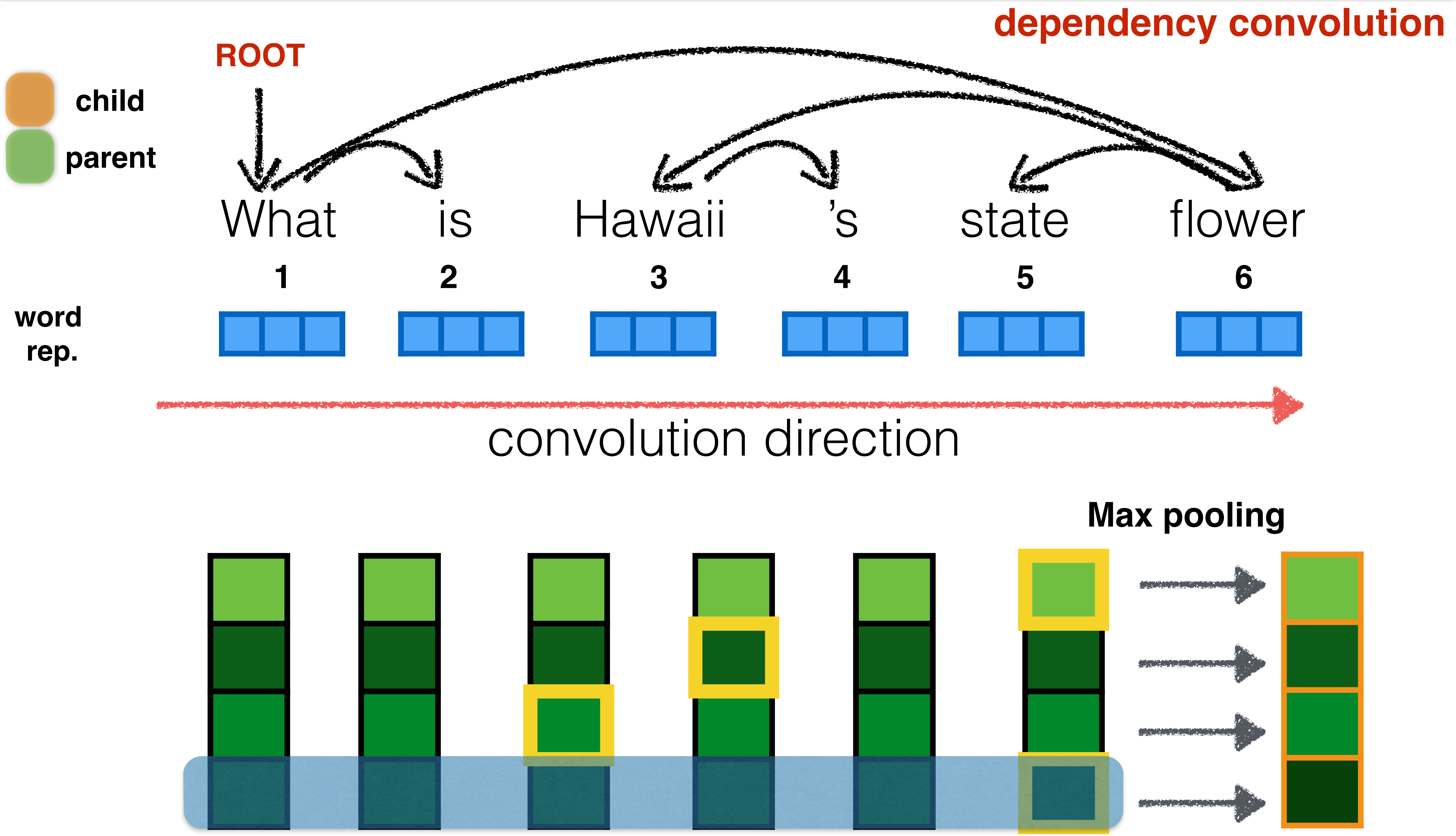
Convolution on Tree



Convolution on Tree

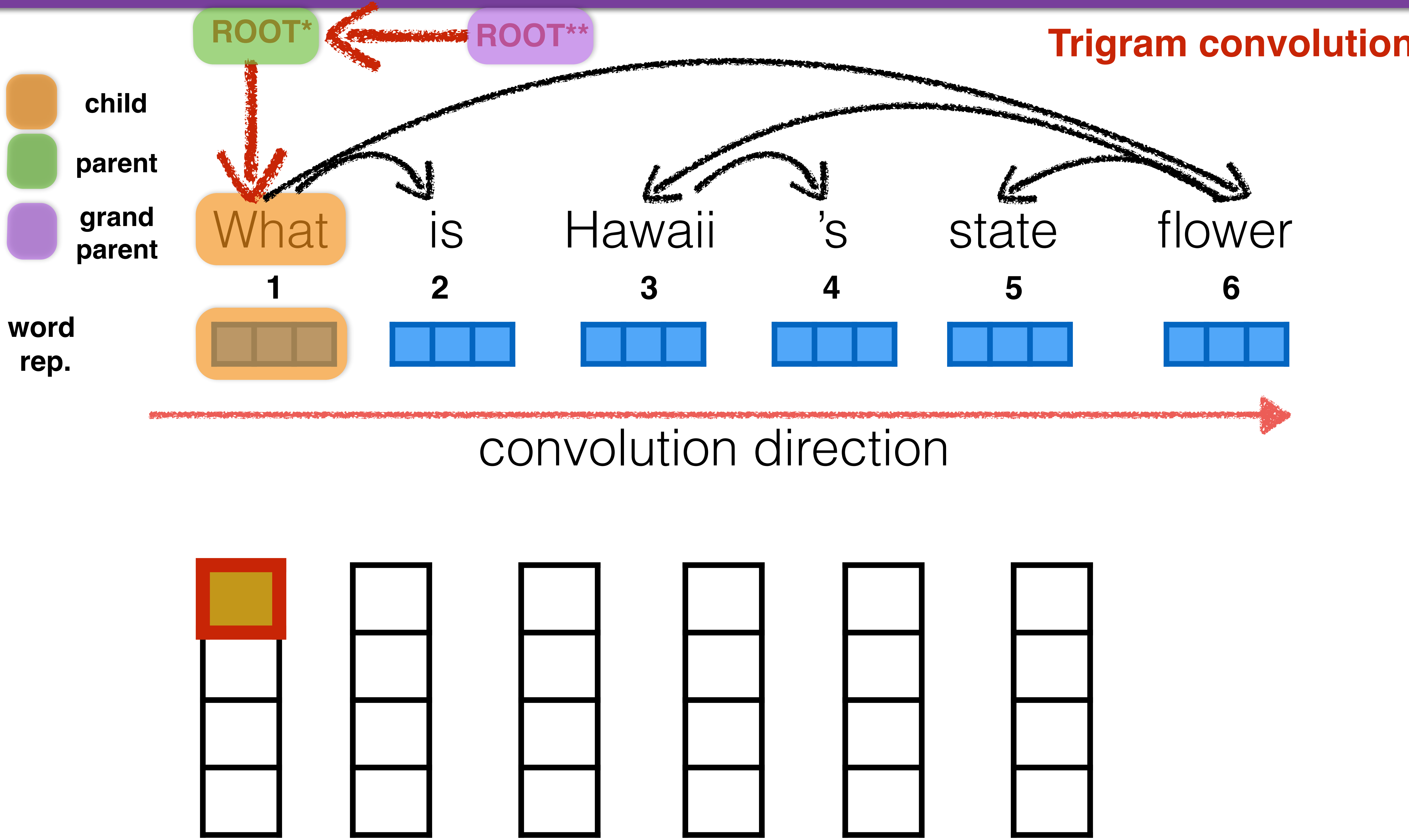


Convolution on Tree

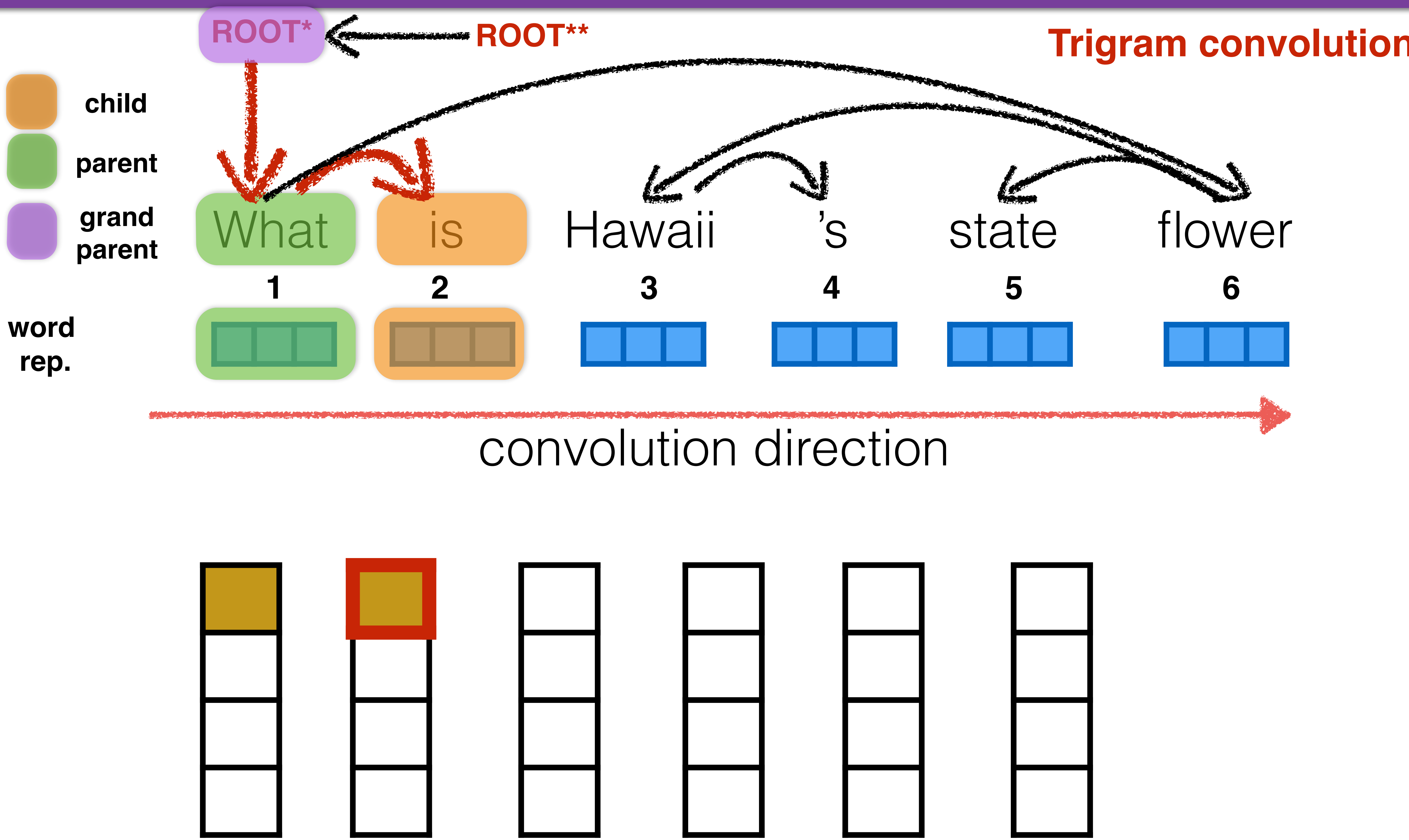


Trigram Convolution on Trees

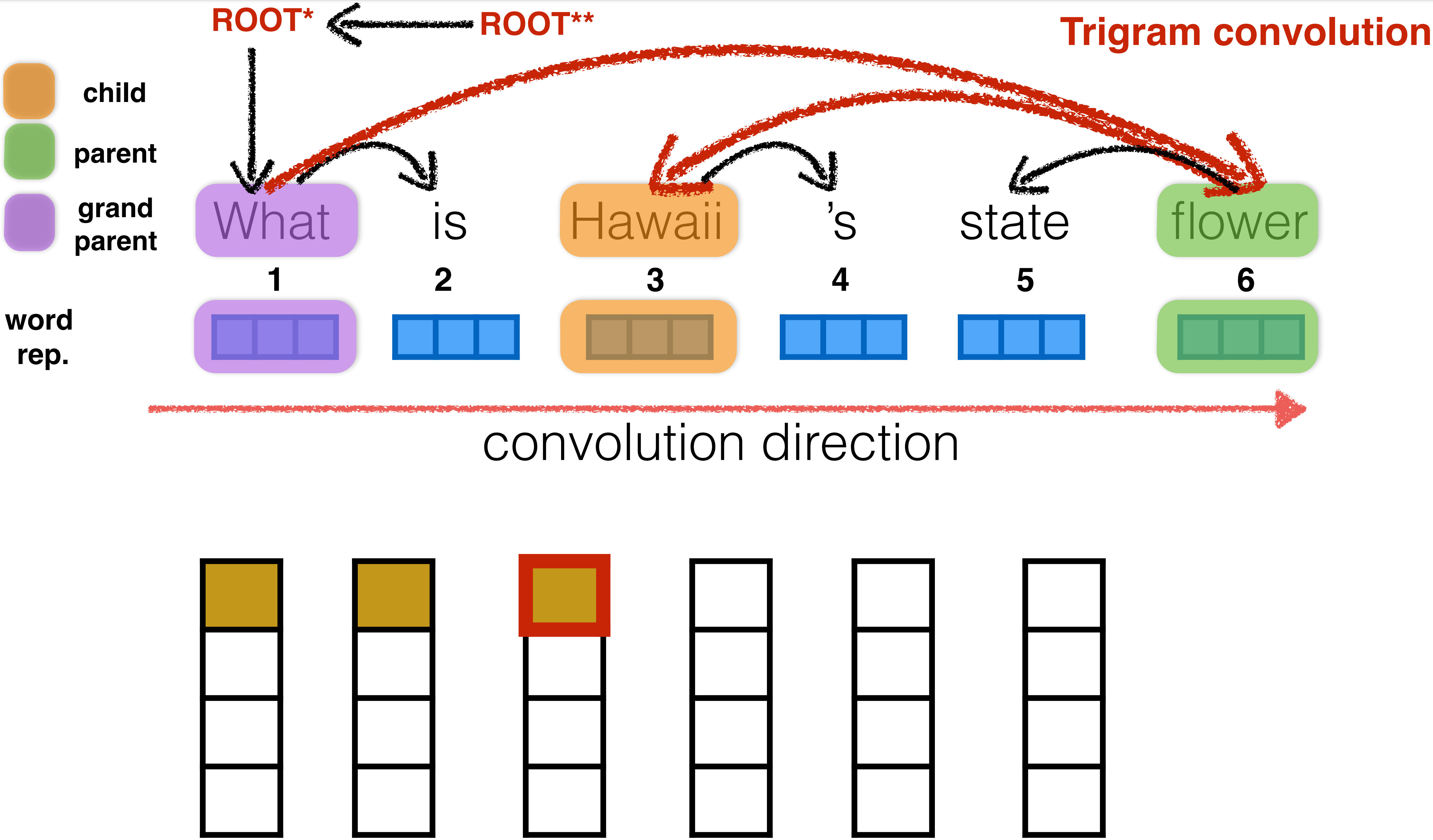
Convolution on Tree



Convolution on Tree

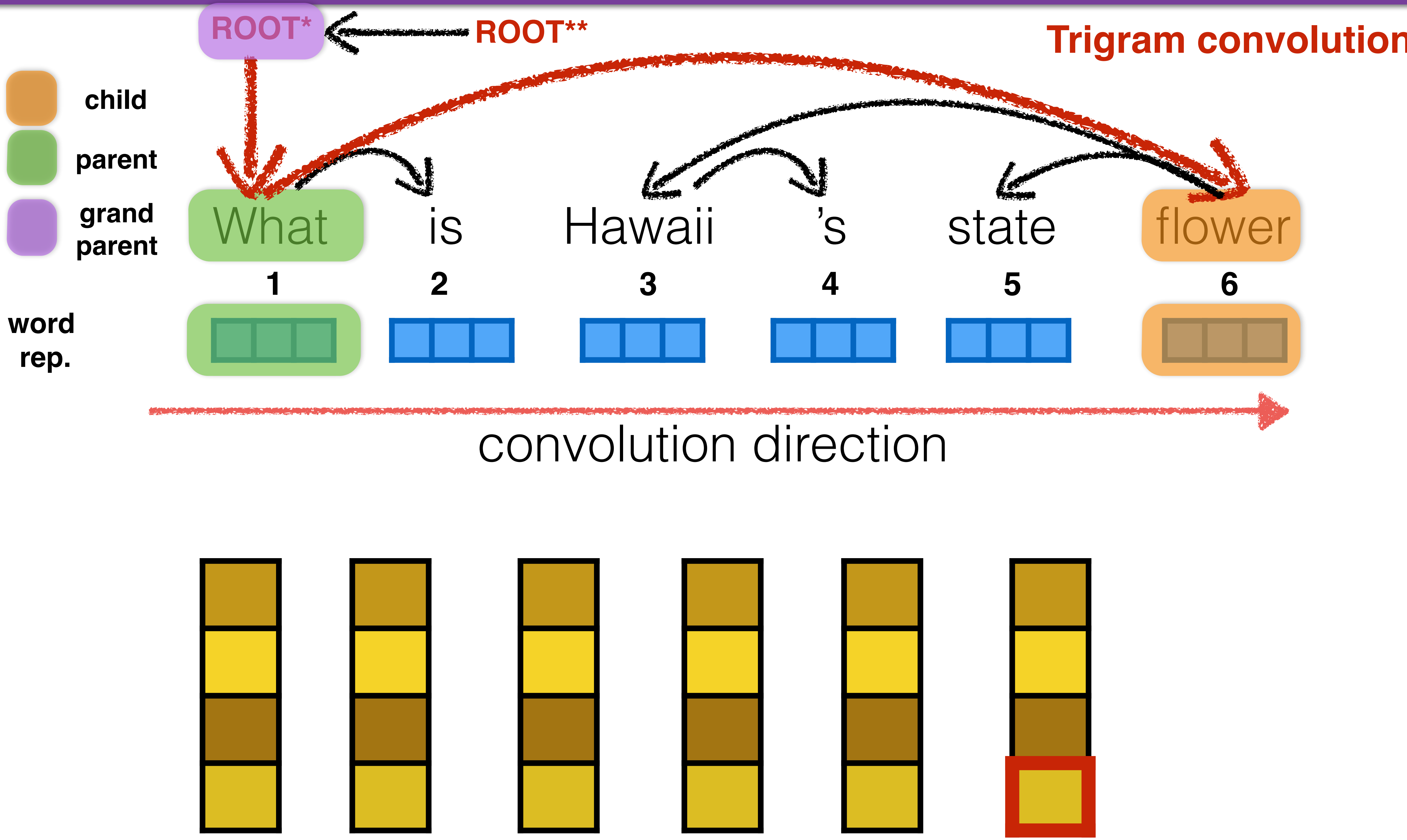


Convolution on Tree



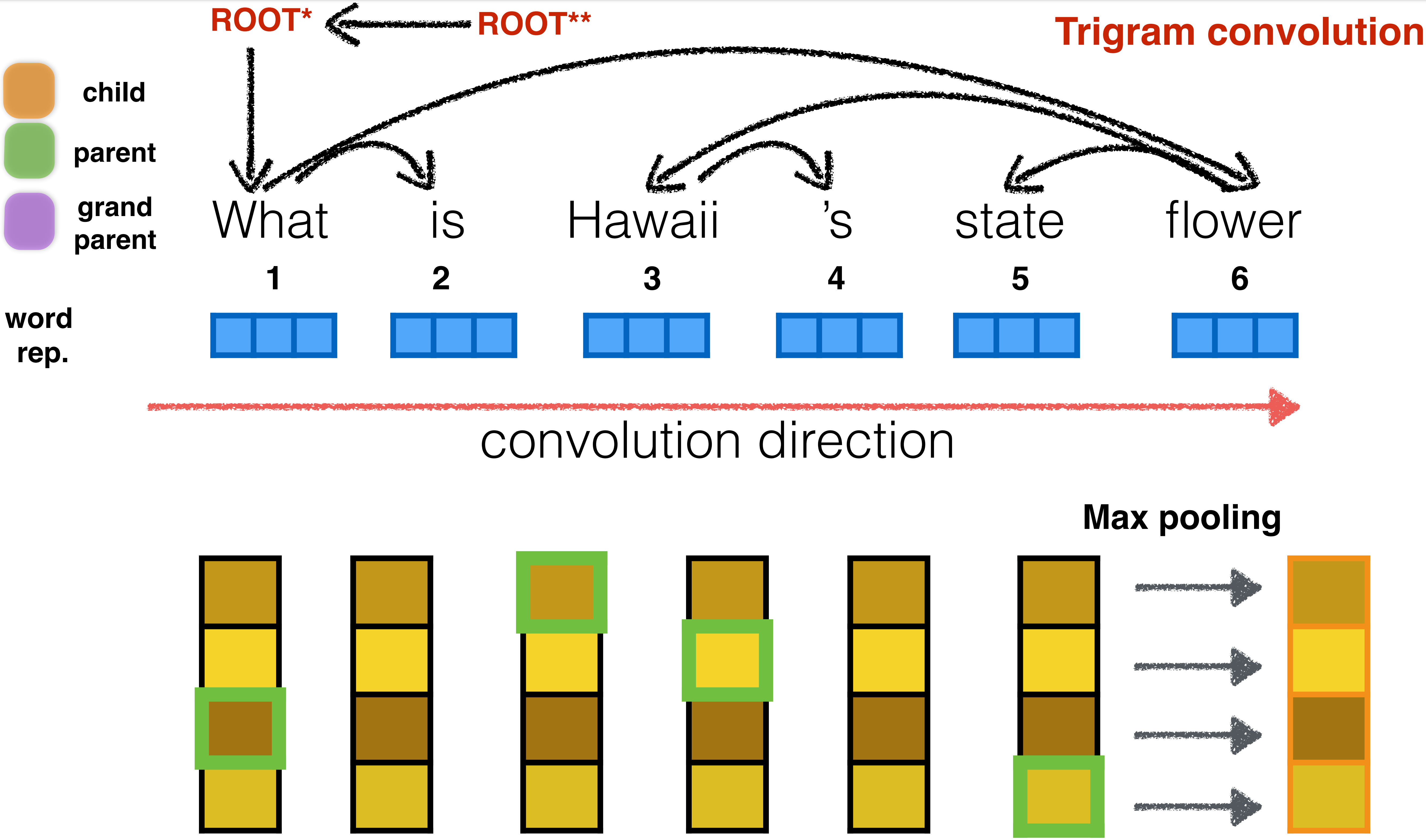
follow the same steps as before…

Convolution on Tree

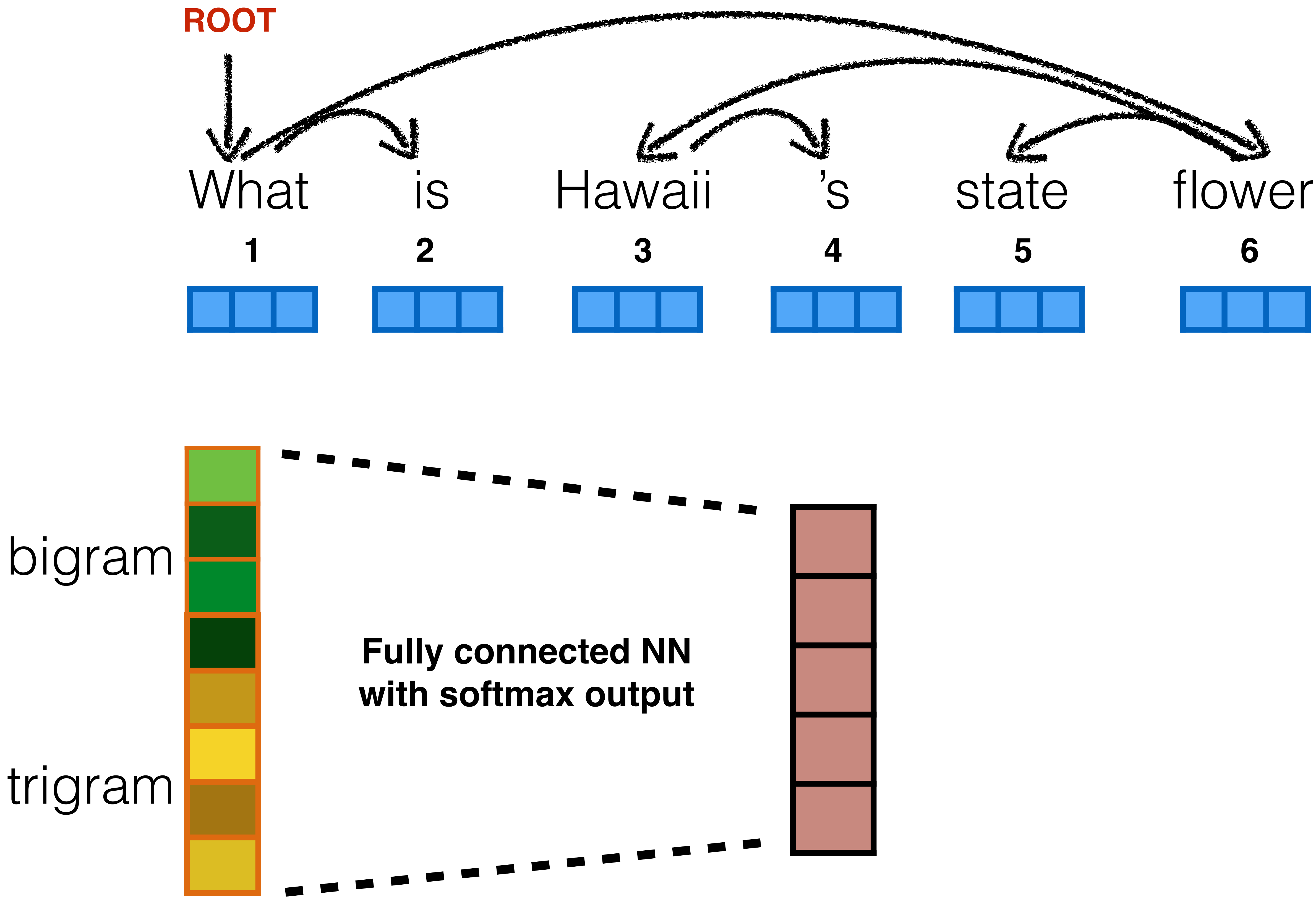


more important words are convolved more often!

Convolution on Tree

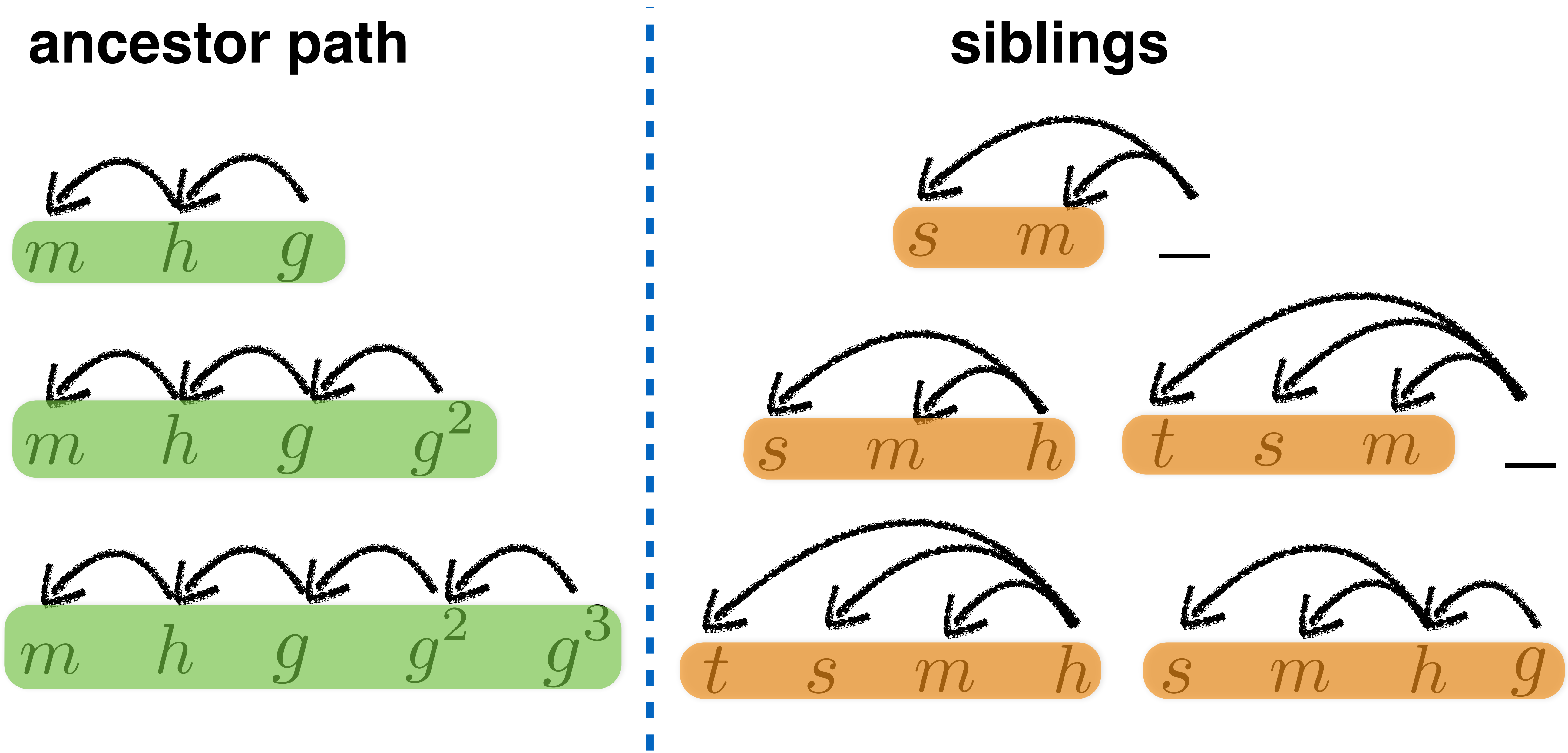


Convolution on Tree



Convolution on Siblings

Besides convolution on ancestor path, we also can capture conjunction information from siblings



Experiments

Tasks:

- Sentimental analysis
- Question classification

Datasets:

Tasks	Dataset	# Classes	Size	Testset
Sentimental Analysis	MR	2	10662	10-CV
	SST1	5	11855	2210
Question Classification	TREC	6	5952	500
	TREC-2	50	5952	500

Sentimental Analysis Data Examples

Sentimental analysis from Rotten Tomatoes (MR & SST-1)

straightforward statements:

simplistic, silly and tedious — Negative

subtle statements:

the film tunes into a grief that could lead a man across centuries — Positive

sentences with adversative:

not for everyone, but for those with whom it will connect, it's a nice departure from standard moviegoing fare — Positive

Sentimental Analysis Experiments Results

Category	Model	MR	SST-1
This work	ancestor	80.4	47.7
	ancestor+sibling	81.7	48.3
	ancestor+sibling+sequential	**81.9**	49.5
CNNs	CNNs-non-static (Kim ’14) — baseline	81.5	48.0
	CNNs-multichannel (Kim ’14)	81.1	47.4
	Deep CNNs (Kalchbrenner+ ’14)	-	48.5
Recursive NNs	Recursive Autoencoder (Socher+ ’11)	77.7	43.2
	Recursive Neural Tensor (Socher+ ’13)	-	45.7
	Deep Recursive NNs (Irsoy+ ’14)	-	**49.8**
Recurrent NNs	LSTM on tree (Zhu+ ’15)	**81.9**	48.0
Other	Paragraph-Vec (Le+ ’14)	-	48.7

Question Classification Examples

Sentence	*Top-level* (TREC)	*Fine-grained* (TREC-2)
How did serfdom develop in and then leave Russia?	DESC	manner
What is Hawaii 's state flower ?	ENTY	plant
What sprawling U.S. state boasts the most airports ?	LOC	state
When was Algeria colonized ?	NUM	date
What person 's head is on a dime ?	HUM	ind
What does the technical term ISDN mean ?	ABBR	exp

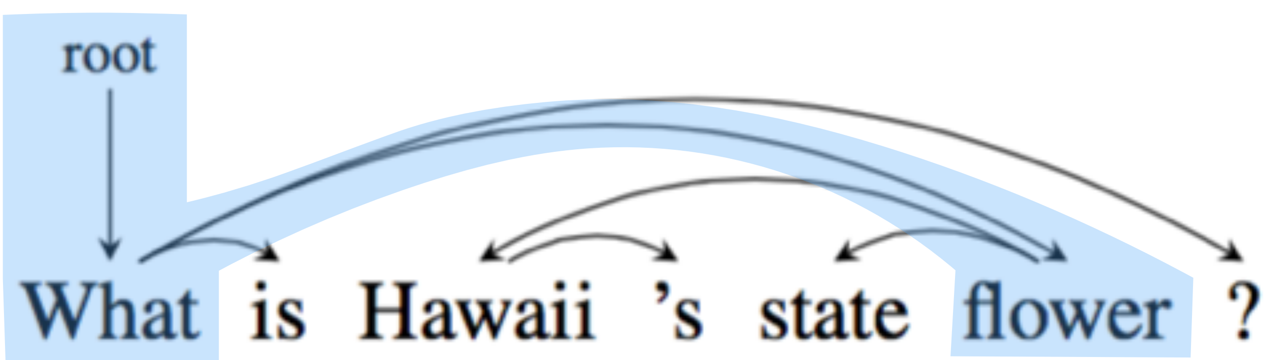
Question Classification Experiments Results

Category	Model	TREC	TREC2
This work	ancestor	95.4	88.4
	ancestor+sibling	**95.6**	89.0
	ancestor+sibling+sequential	95.4	88.8
CNNs	CNNs-non-static (Kim ’14) — baseline	93.6	86.4
	CNNs-multichannel (Kim ’14)	92.2	86.0
	Deep CNNs (Kalchbrenner+ ’14)	93.0	-
Hand-coded	SVMs (Silva+ ’11)*	95.0	**90.8**

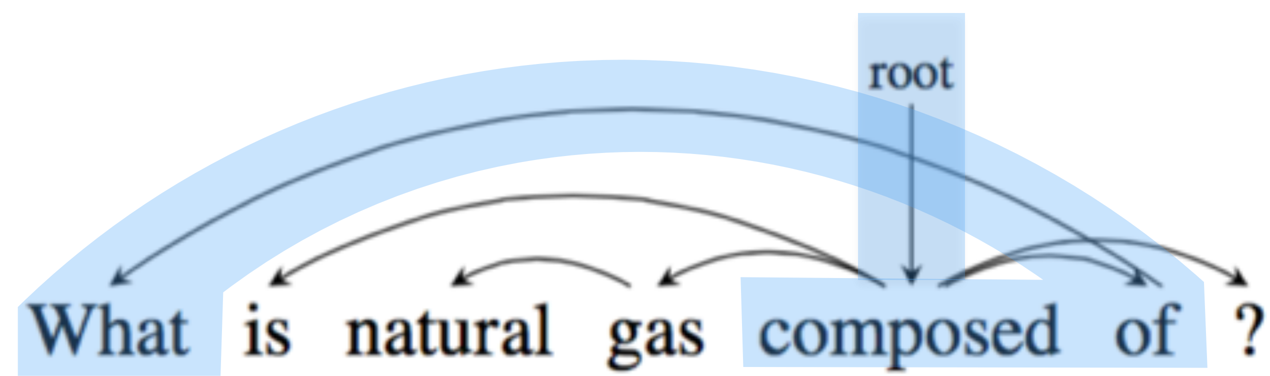
we achieved the highest published accuracy on TREC.

Error Analysis :-)

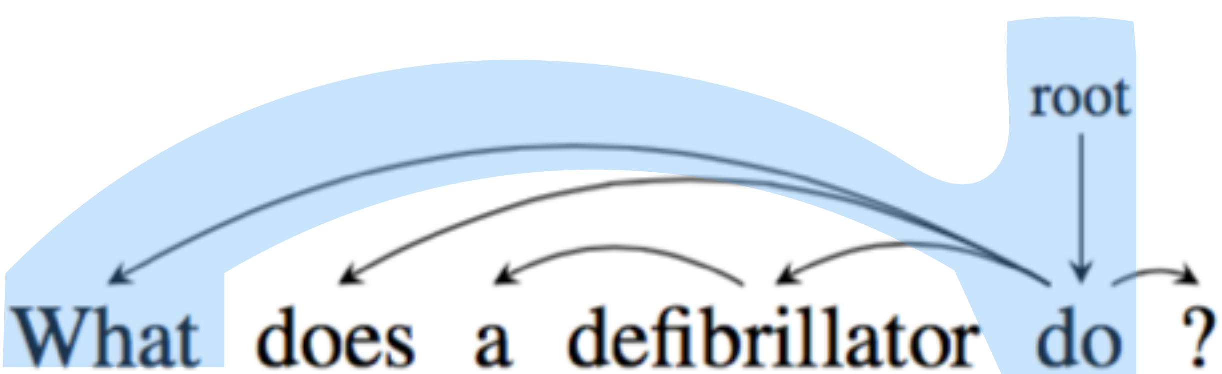
Cases which we do better than Baseline:



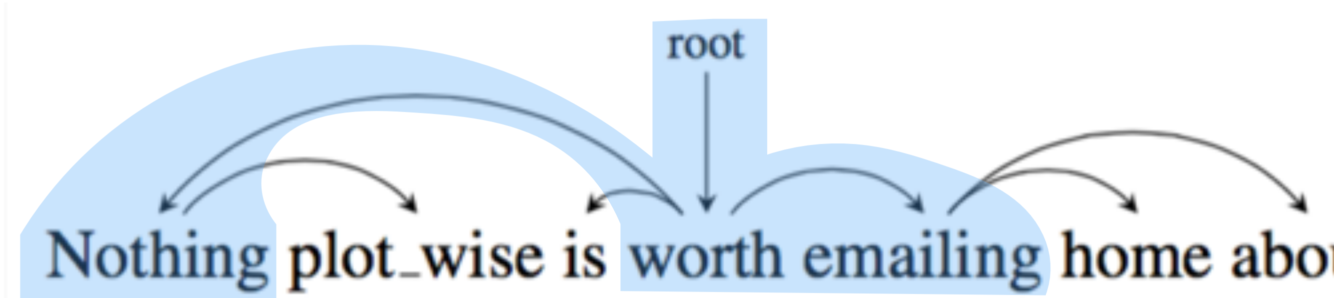
Gold/Ours: Enty Baseline: Loc



Gold/Ours: Enty Baseline: Desc



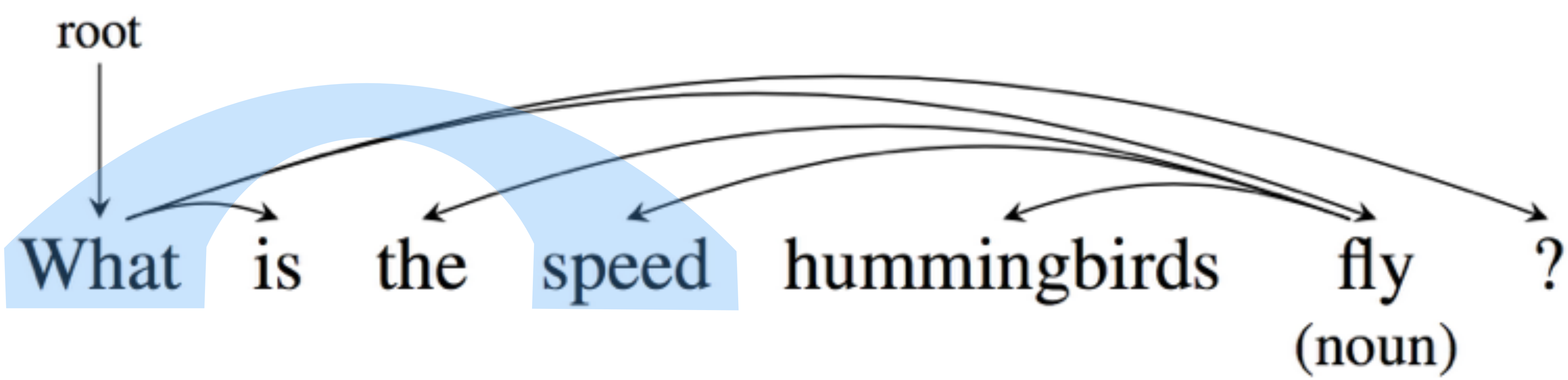
Gold/Ours: Desc Baseline: Enty



Gold/Ours: Mild Neg Baseline: Mild Pos

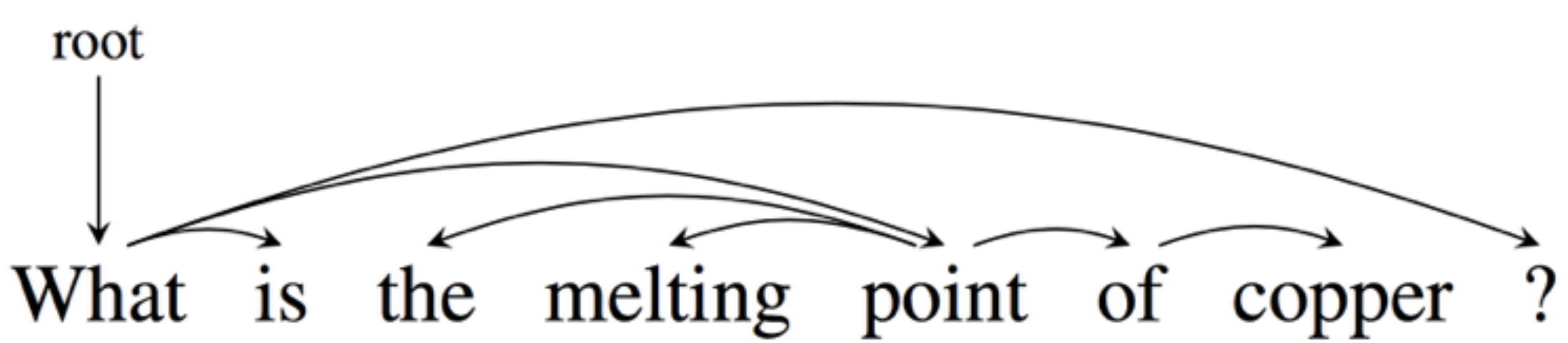
Error Analysis :-(

Cases which we make mistakes:



Gold: Num   Ours: Enty   Baseline: Num

Cases which we and baseline make mistakes:



Gold: Num   Ours: Enty   Baseline: Desc

http://cogcomp.cs.illinois.edu/Data/QA/QC/definition.html

Conclusions

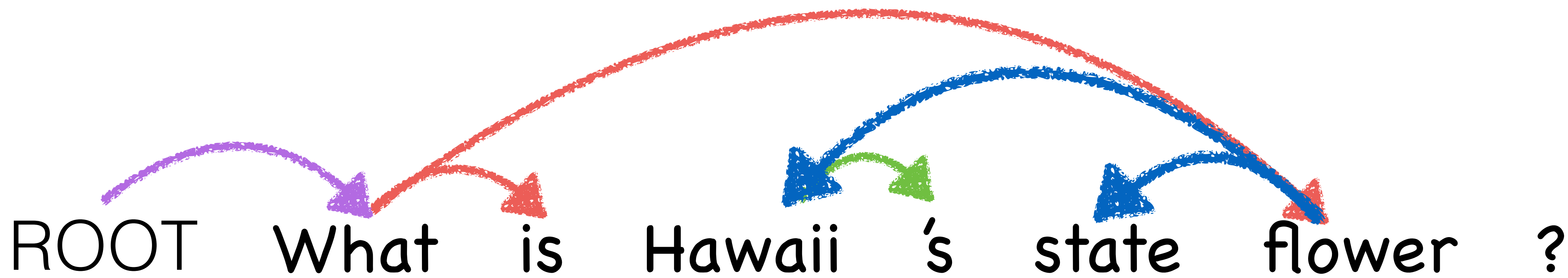
Pros:

- Dependency-based convolution captures long-distance information.
- It outperforms sequential CNN in all four datasets.
  - highest published accuracy on TREC.

Cons:

- Our model's accuracy depends on parser quality.

Deep Learning can and should be combined with linguistic intuitions.



Thank you !